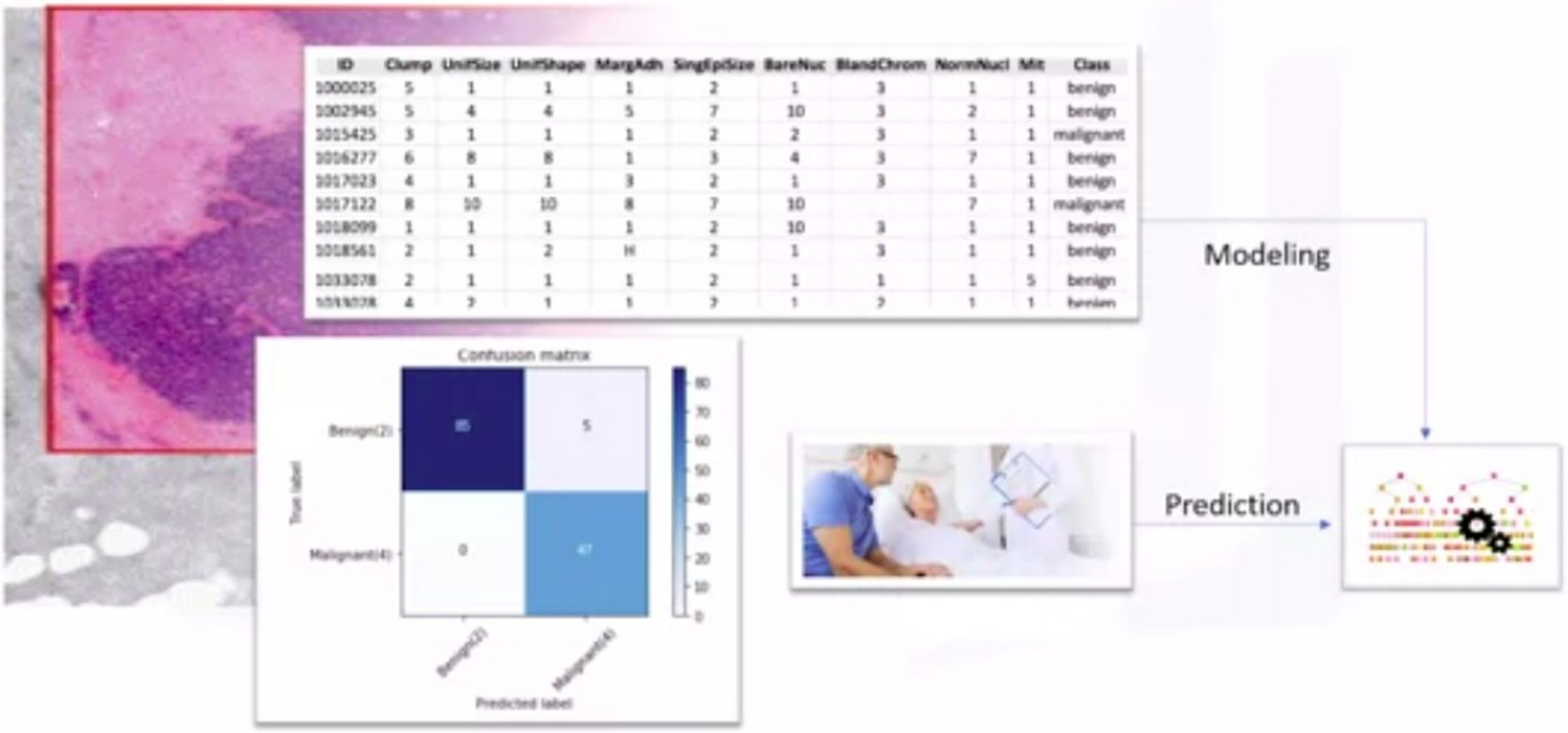
Week 1 to 6

Machine Learning with Python

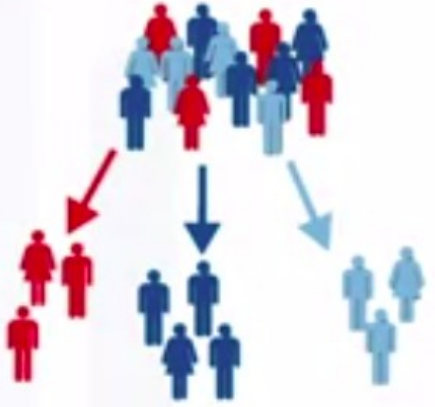
Welcome

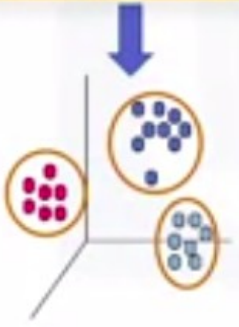
Use ML to make predictions …

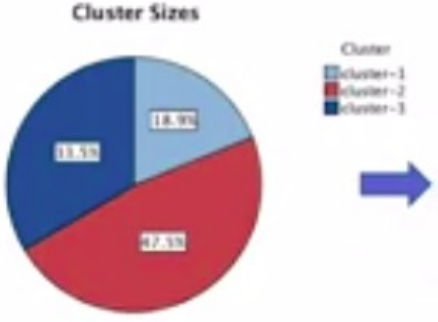


Use ML for customer segmentation

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1



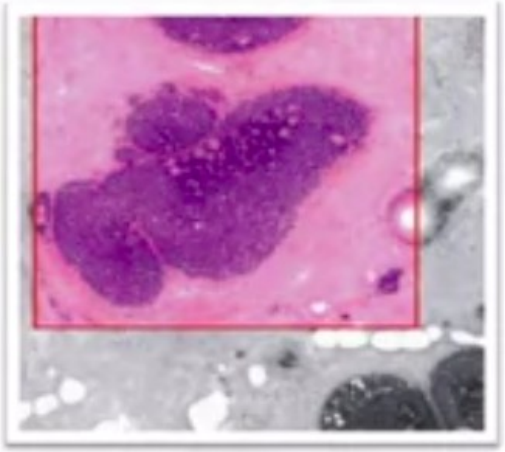


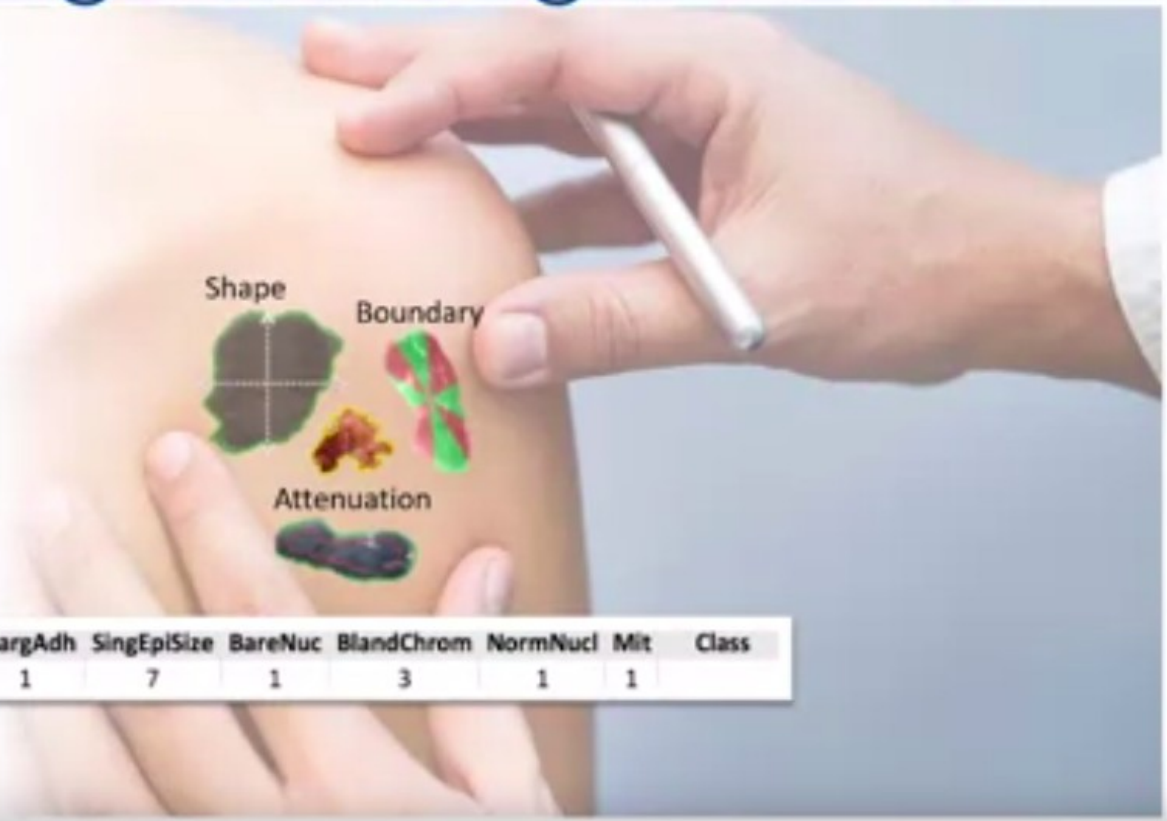


Cluster	Segment Name
cluster-1	AFFULUENT AND MIDDLE AGED
cluster-2	YOUNG EDUCATED AND MIDDLE INCOME
cluster-3	YOUNG AND LOW INCOME


Introduction to Machine Learning
IBM Developer
SKILLS NETWORK
0:07 / 8:53

Is this a benign or malignant cell?





ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000015	6	1	1	1	7	1	3	1	1	

Machine learning helps with predictions!

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

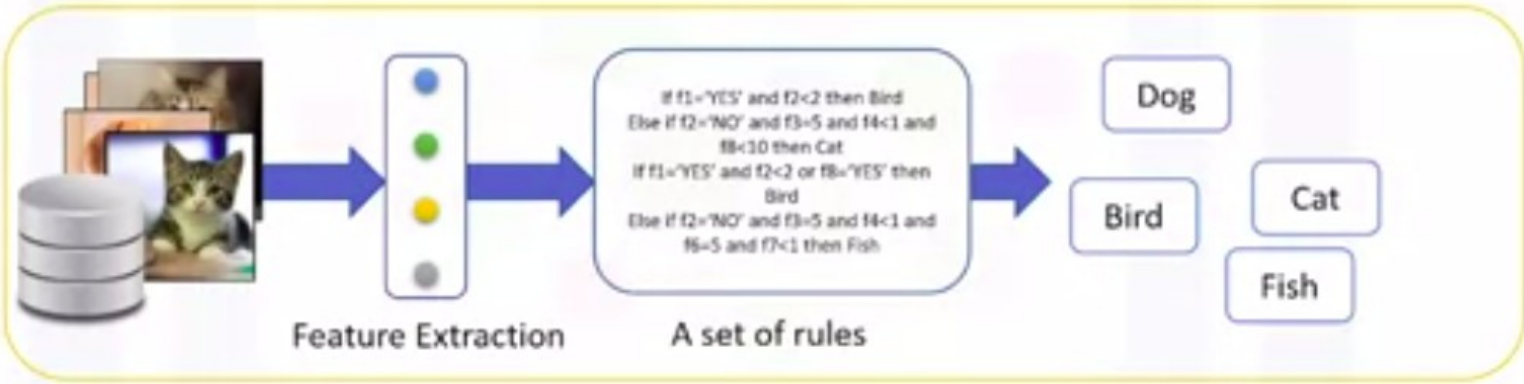
Machine learning helps with predictions!

What is machine learning?

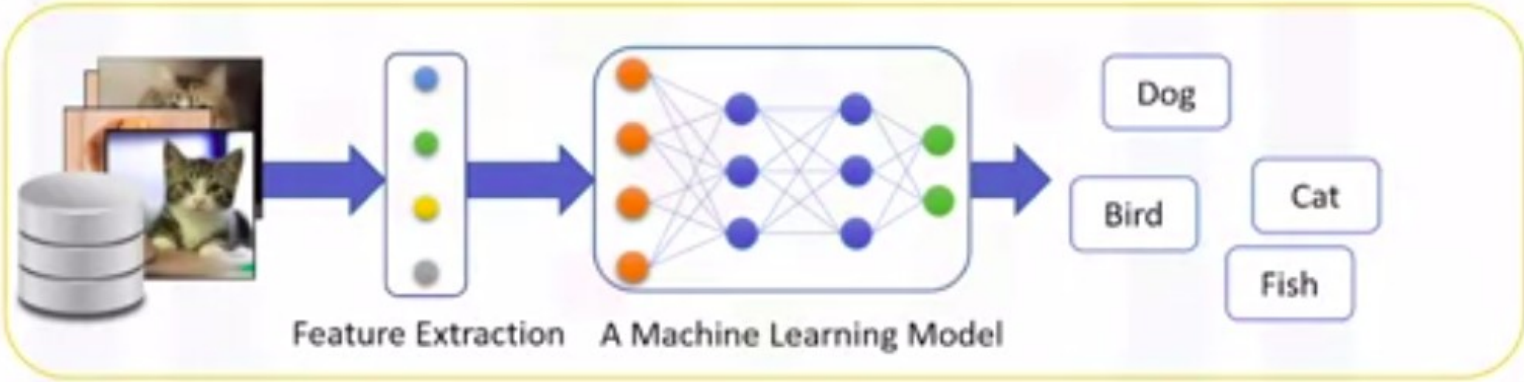
Machine learning is the subfield of computer science that gives "**computers the ability to learn without being explicitly programmed.**"

Arthur Samuel
American pioneer in the field of computer gaming and artificial intelligence, coined the term "machine learning" in 1959 while at IBM.

How machine learning works?



How machine learning works?



Major machine learning techniques

- Regression/Estimation
  - Predicting continuous values
- Classification
  - Predicting the item class/category of a case
- Clustering
  - Finding the structure of data; summarization
- Associations
  - Associating frequent co-occurring items/events

Major machine learning techniques

- Anomaly detection
  - Discovering abnormal and unusual cases
- Sequence mining
  - Predicting next events; click-stream (Markov Model, HMM)
- Dimension Reduction
  - Reducing the size of data (PCA)
- Recommendation systems
  - Recommending items

Question

Which Machine Learning technique is proper for grouping of similar cases in a dataset, for example to find similar patients, or for customers
segmentation in a bank?

- ○ Classification
- ◉ Clustering
- ○ Regression
- ○ Recommender systems

✓ **Correct**

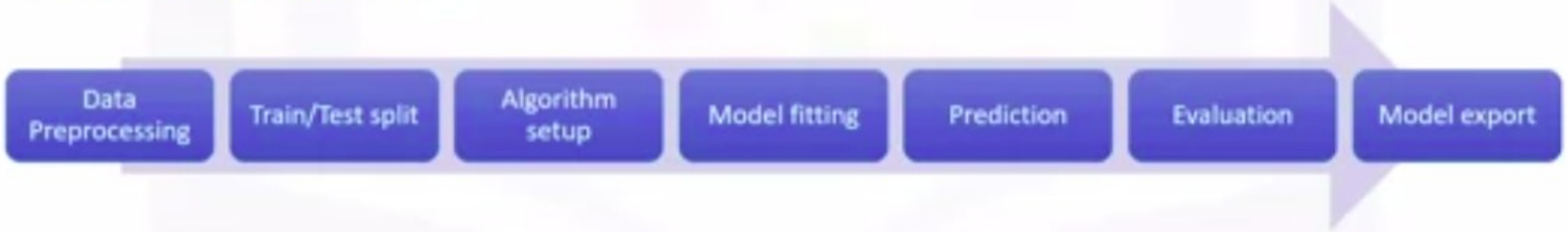
Skip Continue

Python for Machine Learning

Python libraries for machine learning

More about scikit-learn

- Free software machine learning library
- Classification, Regression and Clustering algorithms
- Works with NumPy and SciPy
- Great documentation
- Easy to implement



Question

Why Scikit is a proper library for Machine Learning (select all the options that are correct)?

- [x] Scikit-learn is a free machine learning library that works with Numpy and Scipy.

Correct

- [x] Scikit-learn has most of machine learning algorithms.

Correct

- [ ] Scikit-learn support all data science languages such as Python, R and Java.

Skip Continue

scikit-learn functions

```
from sklearn import preprocessing
X = preprocessing.StandardScaler().fit(X).transform(X)
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)
```

```
from sklearn import svm
clf = svm.SVC(gamma=0.001, C=100.)
```

```
clf.fit(X_train, y_train)
```

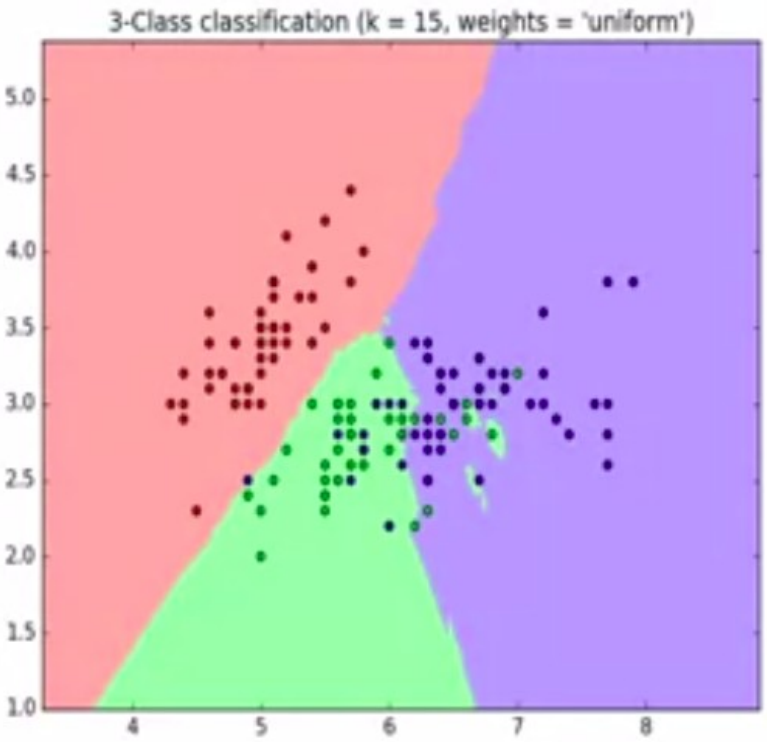
```
clf.predict(X_test)
```

```
from sklearn.metrics import confusion_matrix
print(confusion_matrix(y_test, yhat, labels=[1,0]))
```

```
import pickle
s = pickle.dumps(clf)
```


Supervised vs Unsupervised
IBM Developer
SKILLS NETWORK
0:01 / 6:04

What is supervised learning?



We "teach the model," then with that knowledge, it can predict unknown or future instances.

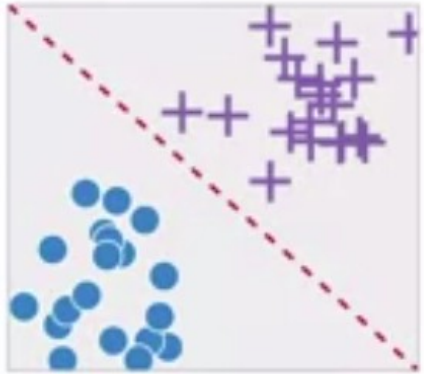
Teaching the model with labeled data

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

What is classification?

Classification is the process of predicting discrete class labels or categories.

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

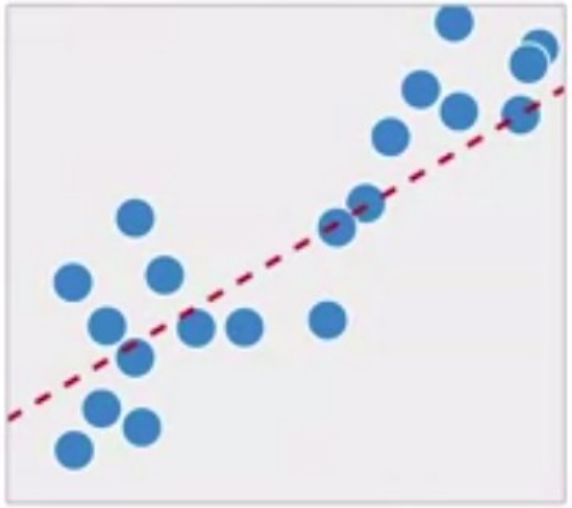


What is regression?

Regression is the process of predicting continuous values.

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Continuous Values



Question

Which technique/s is/are considered as Supervised learning?

- [ ] Clustering
- [x] Regression

Correct

- [x] Classification

Correct

Skip Continue

What is unsupervised learning?

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio
1	41	2	6	19	0.124	1.073	NBA001	6.3
2	47	1	26	100	4.582	8.218	NBA021	12.8
3	33	2	10	57	6.111	5.802	NBA013	20.9
4	29	2	4	19	0.681	0.516	NBA009	6.3
5	47	1	31	253	9.308	8.908	NBA008	7.2
6	40	1	23	81	0.998	7.831	NBA016	10.9
7	38	2	4	56	0.442	0.454	NBA013	1.6
8	42	3	0	64	0.279	3.945	NBA009	6.6
9	26	1	5	18	0.575	2.215	NBA006	15.5
10	47	3	23	115	0.653	3.947	NBA011	4
11	44	3	8	88	0.285	5.083	NBA010	6.1
12	34	2	9	40	0.374	0.266	NBA003	1.6

Unsupervised learning techniques:

- Dimension reduction
- Density estimation
- Market basket analysis
- Clustering

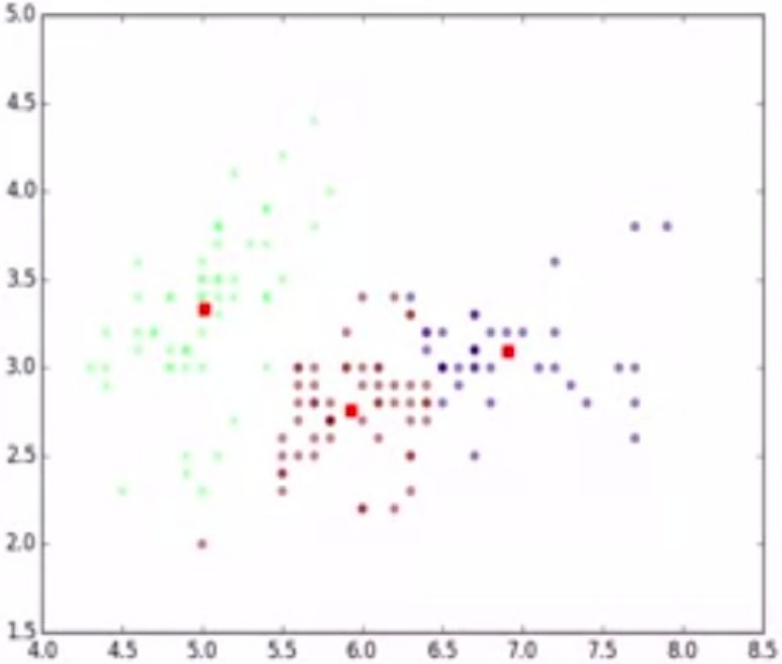
ALL OF THIS DATA IS UNLABELED

The model works on its own to discover information.

What is clustering?

Clustering is grouping of data points or objects that are somehow similar by:

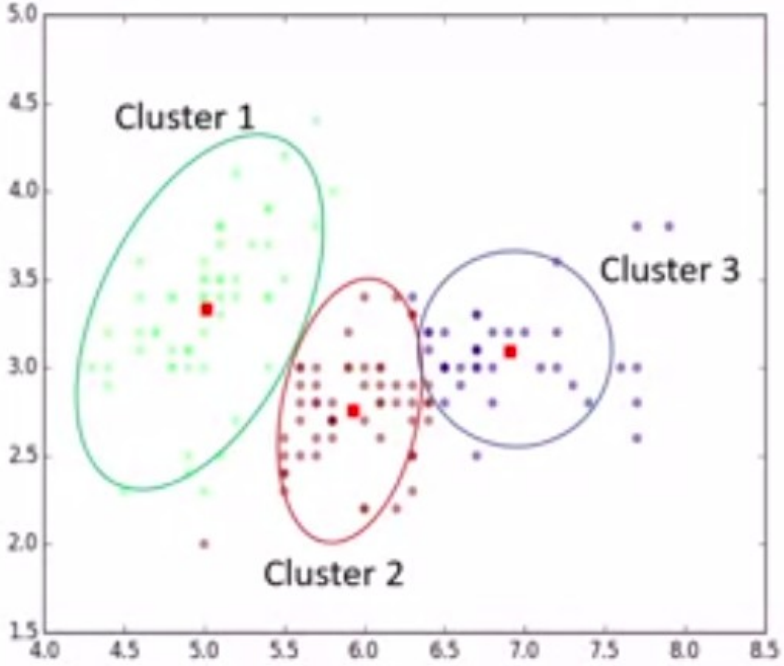
- Discovering structure
- Summarization
- Anomaly detection



What is clustering?

Clustering is grouping of data points or objects that are somehow similar by:

- Discovering structure
- Summarization
- Anomaly detection



Supervised vs unsupervised learning

Supervised Learning

- **Classification:**
  Classifies labeled data
- **Regression:**
  Predicts trends using previous labeled data
- Has more evaluation methods than unsupervised learning
- Controlled environment

Unsupervised Learning

- **Clustering:**
  Finds patterns and groupings from unlabeled data
- Has fewer evaluation methods than supervised learning
- Less controlled environment

Back

Practice Quiz: Intro to Machine Learning

Practice Quiz • 10 min • 3 total points

Congratulations! You passed!

Grade received 100% **To pass** 66% or higher

Go to next item

1. Supervised learning deals with unlabeled data, while unsupervised learning deals with labelled data. **1 / 1 point**

- ○ True
- ◉ False

Correct
Correct! Unsupervised learning deals with unlabeled data, and supervised learning deals with labelled data

2. The "Regression" technique in Machine Learning is a group of algorithms that are used for: **1 / 1 point**

Practice Quiz: Intro to Machine Learning

Practice Quiz • 10 min • 3 total points

Correct! Unsupervised learning deals with unlabeled data, and supervised learning deals with labelled data

2. The "Regression" technique in Machine Learning is a group of algorithms that are used for: **1 / 1 point**

- ○ Prediction of class/category of a case; for example, a cell is benign or malignant, or a customer will churn or not.
- ○ Finding items/events that often co-occur; for example grocery items that are usually bought together by a customer.
- ◉ Predicting a continuous value; for example predicting the price of a house based on its characteristics.

Correct
Correct! Regression techniques are used for continuous variable prediction, whereas classification techniques handle dependent variables with discrete classes.

3. When comparing Supervised with Unsupervised learning, is this sentence True or False? **1 / 1 point**

Correct
Correct! Regression techniques are used for continuous variable prediction, whereas classification techniques handle dependent variables with discrete classes.

3. When comparing Supervised with Unsupervised learning, is this sentence True or False? **1 / 1 point**

In contrast to Supervised learning, Unsupervised learning has more models and more evaluation methods that can be used in order to ensure the outcome of the model is accurate.

- ○ True
- ◉ False

Correct
Correct! Unsupervised learning has fewer models and evaluation methods than Supervised learning.

What is regression?

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Regression is the process of predicting a continuous value

What is regression?

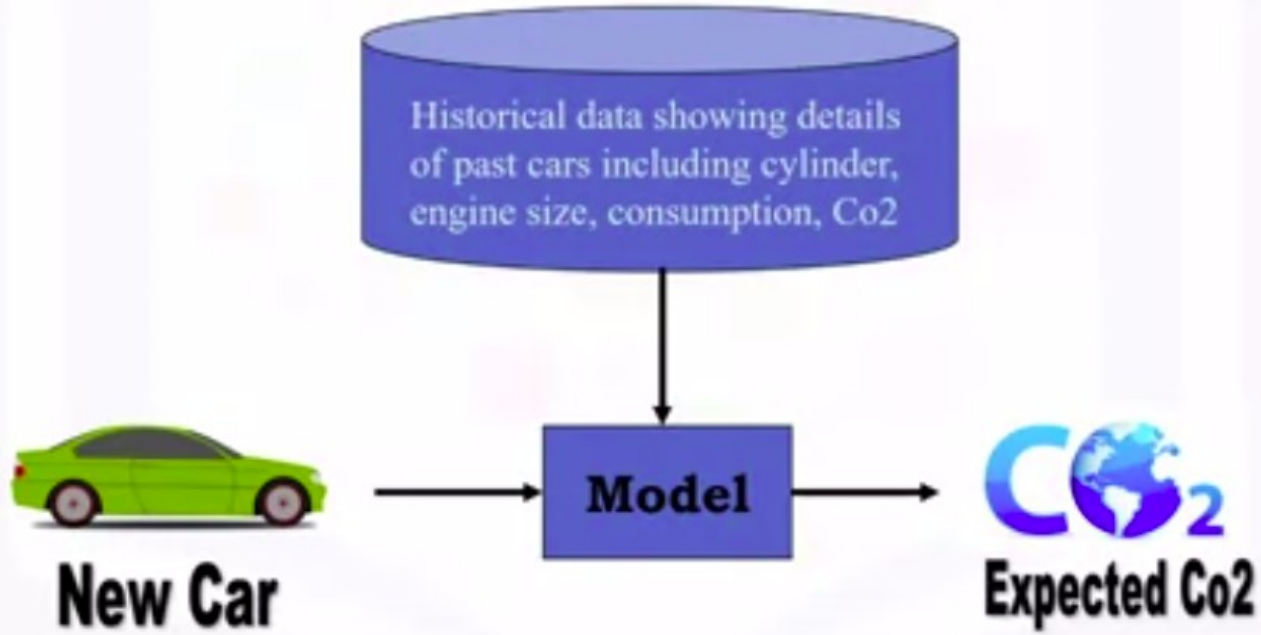
X: Independent variable

Y: Dependent variable

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Regression is the process of predicting a continuous value

What is a regression model?



Types of regression models

- Simple Regression:
  - Simple Linear Regression
  - Simple Non-linear Regression

Predict co2emission vs EngineSize of all cars

- Multiple Regression:
  - Multiple Linear Regression
  - Multiple Non-linear Regression

Predict co2emission vs EngineSize and Cylinders of all cars

Question

Which one is a sample application of regression?

- ○ Predicting whether a patient has cancer or not.
- ○ Grouping of similar houses in an area.
- ◉ Forecasting rainfall amount for next day.
- ○ Predicting if a team will win or not.

✓ **Correct**

Skip Continue

Applications of regression

- Sales forecasting
- Satisfaction analysis
- Price estimation
- Employment income

Regression algorithms

- Ordinal regression
- Poisson regression
- Fast forest quantile regression
- Linear, Polynomial, Lasso, Stepwise, Ridge regression
- Bayesian linear regression
- Neural network regression
- Decision forest regression
- Boosted decision tree regression
- KNN (K-nearest neighbors)

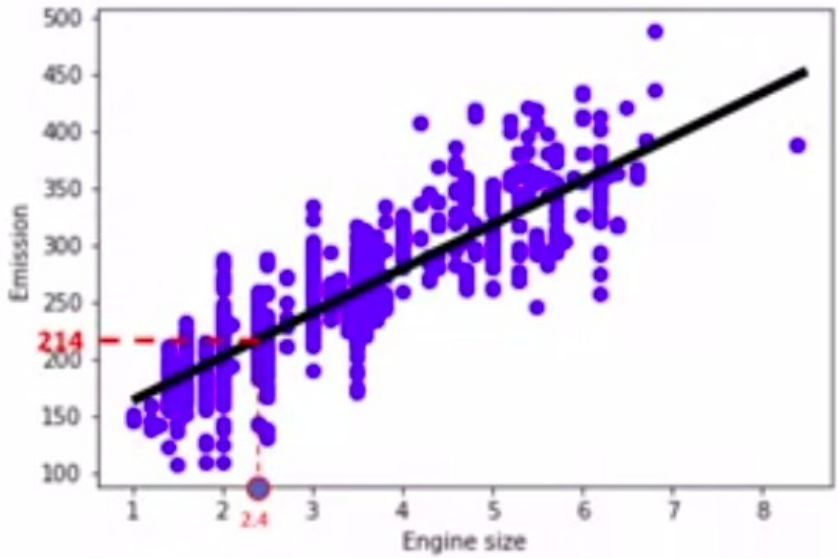
Simple Linear Regression

Linear regression topology

- Simple Linear Regression:
  - Predict co2emission vs EngineSize of all cars
    - Independent variable (x): EngineSize
    - Dependent variable (y): co2emission
- Multiple Linear Regression:
  - Predict co2emission vs EngineSize and Cylinders of all cars
    - Independent variable (x): EngineSize, Cylinders, etc
    - Dependent variable (y): co2emission

How does linear regression work?

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

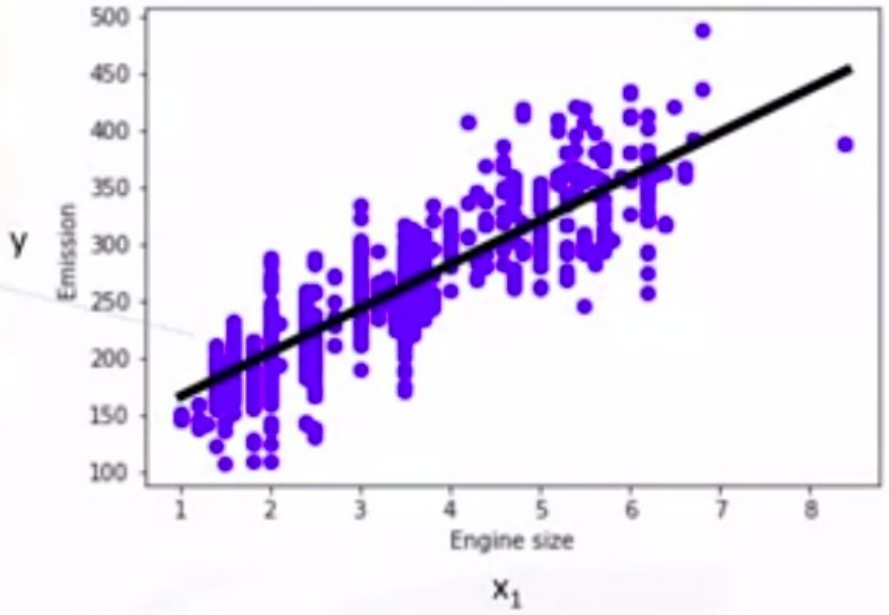


Linear regression model representation

$$\hat{y} = \theta_0 + \theta_1 x_1$$

response variable

a single predictor



How to find the best fit?

$x_1 = 5.4$ independent variable

$y = 250$ actual Co2 emission of x1

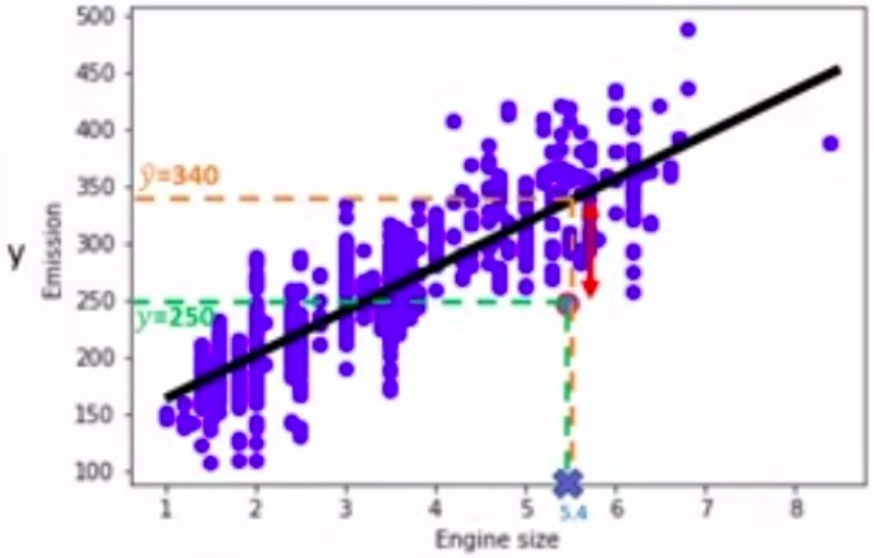
$\hat{y} = \theta_0 + \theta_1 x_1$

$\hat{y} = 340$ the predicted emission of x1

Error = $y - \hat{y}$

= 250 – 340

= -90



How to find the best fit?

$x_1 = 5.4$ independent variable

$y = 250$ actual Co2 emission of x1

$\hat{y} = \theta_0 + \theta_1 x_1$

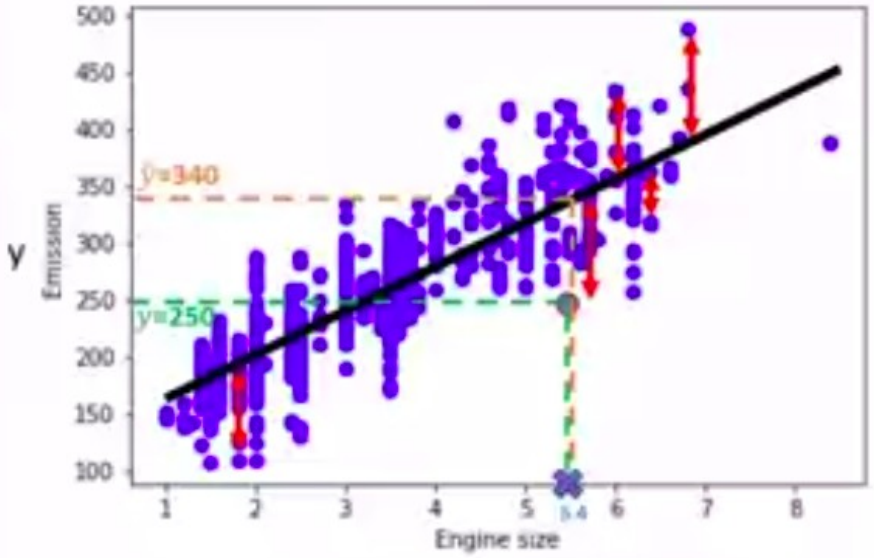
$\hat{y} = 340$ the predicted emission of x1

Error = $y - \hat{y}$

$= 250 - 340$

$= -90$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



Estimating the parameters

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

X_1 y

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$\bar{x} = (2.0 + 2.4 + 1.5 + \dots)/9 = 3.03$

$\bar{y} = (196 + 221 + 136 + \dots)/9 = 226.22$

$$\theta_1 = \frac{(2.0 - 3.03)(196 - 226.22) + (2.4 - 3.03)(221 - 226.22) + \dots}{(2.0 - 3.03)^2 + (2.4 - 3.03)^2 + \dots}$$

$\theta_1 = 39$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

$\theta_0 = 226.22 - 39 * 3.03$

$\theta_0 = 125.74$

$$\hat{y} = 125.74 + 39 x_1$$

Predictions with linear regression

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$Co2Emission = \theta_0 + \theta_1\ EngineSize$$

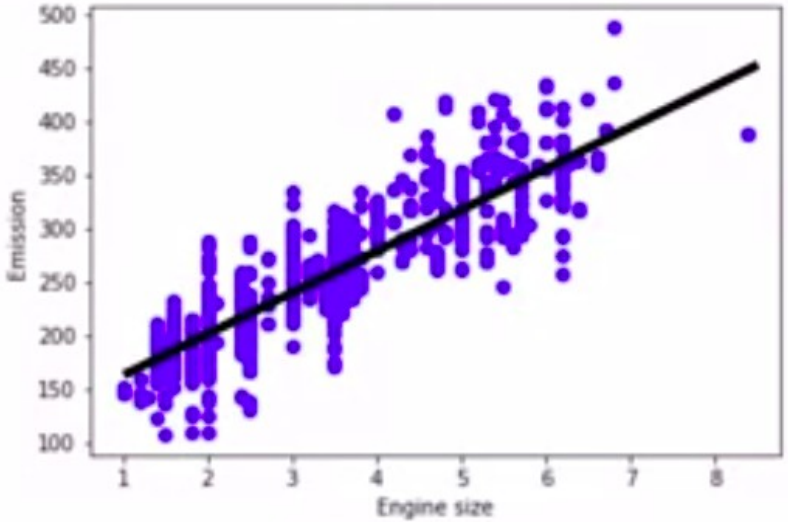
$$Co2Emission = 125 + 39\ EngineSize$$

$$Co2Emission = 125 + 39 \times 2.4$$

$$Co2Emission = 218.6$$

Pros of linear regression

- Very fast
- No parameter tuning
- Easy to understand, and highly interpretable



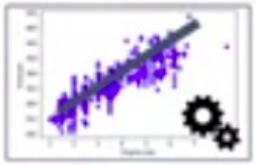

Model Evaluation in
Regression Models
IBM Developer
SKILLS NETWORK
0:05 / 8:32

Best approach for most accurate results?

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	212

Train

Test



	Prediction
6	234
7	256
8	267
9	210

Actual values

Predicted values

Calculating the accuracy of a model

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	212

Test

y

Actual values

$$Error = \frac{(232 - 234) + (255 - 256) + \ldots}{4}$$

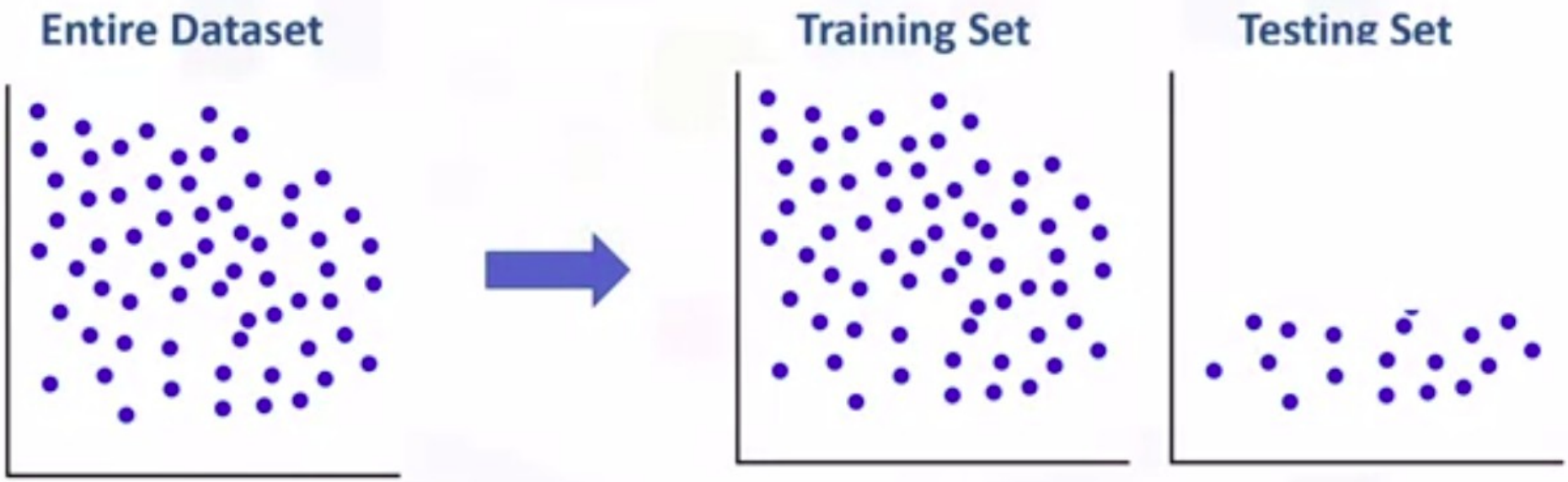
$$Error = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$\hat{y}$

	Prediction
6	234
7	256
8	267
9	210

Predicted values

Train and test on the same dataset



What is training & out-of-sample accuracy?

- **Training Accuracy**
  - High training accuracy isn't necessarily a good thing
  - Result of over-fitting
    - Over-fit: the model is overly trained to the dataset, which may capture noise and produce a non-generalized model

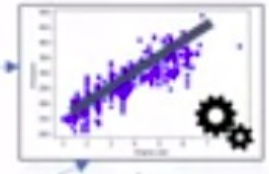
- **Out-of-Sample Accuracy**
  - It's important that our models have a high, out-of-sample accuracy
  - How can we improve out-of-sample accuracy?

Train/Test split evaluation approach

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	212

Train

Test



	Prediction
6	234
7	256
8	267
9	210

Actual values

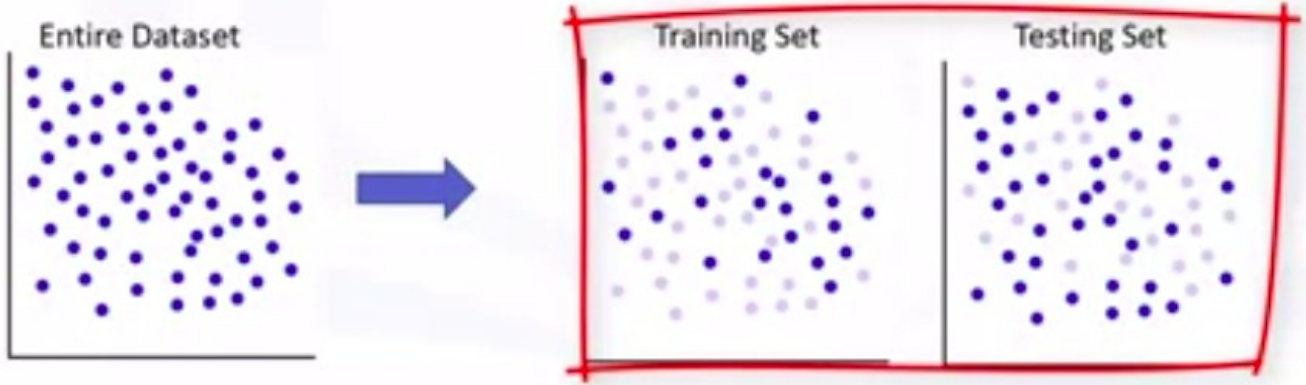
Predicted values

Train/Test split evaluation approach

Test on a portion of train set

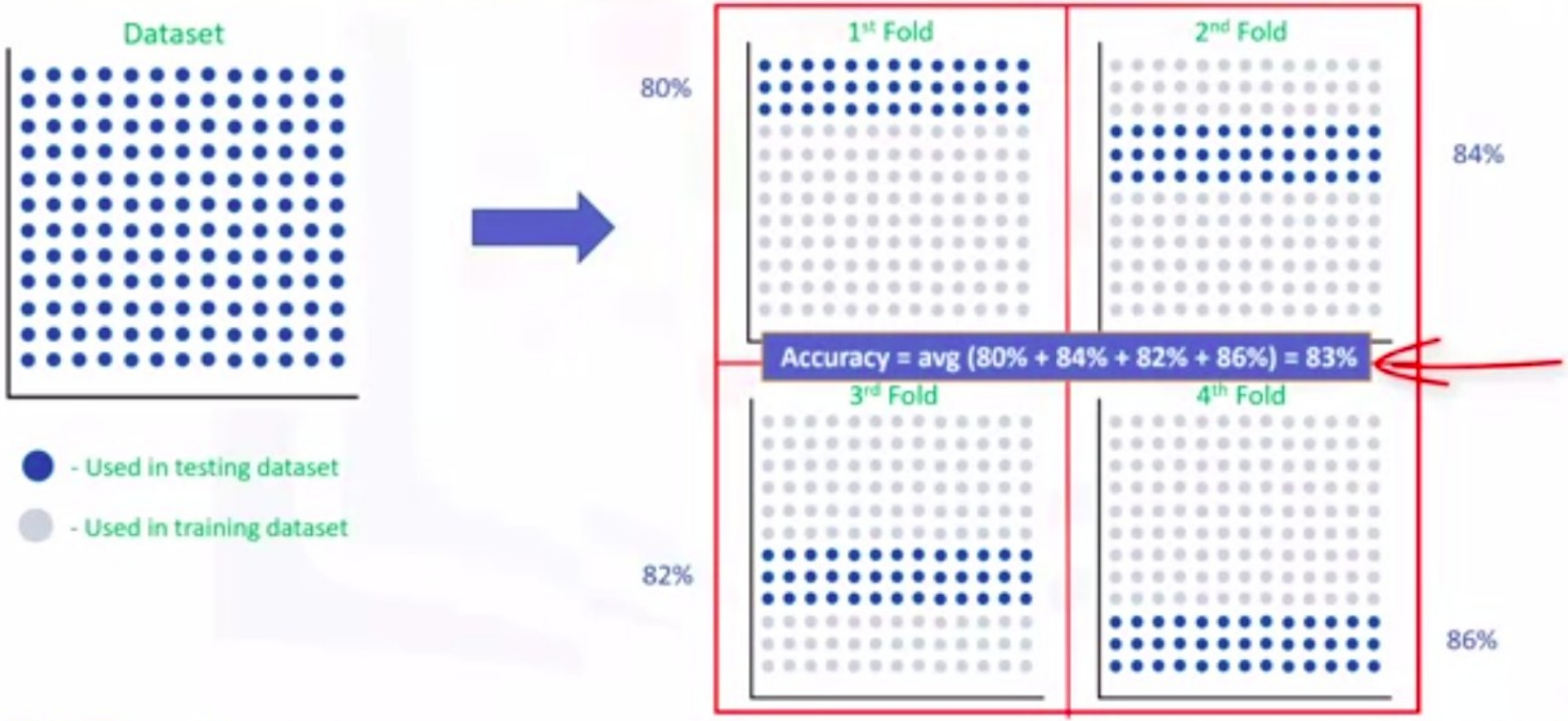
- Test-set is a portion of the train-set
- High "training accuracy"
- Low "out-of-sample accuracy"

Train/Test Split



- Mutually exclusive
- More accurate evaluation on out-of-sample accuracy
- Highly dependent on which datasets the data is trained and tested

How to use K-fold cross-validation?




Evaluation Metrics in
Regression Models
IBM Developer
SKILLS NETWORK
0:08 / 3:11

Regression accuracy

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	212

Test

y

Actual values

$$Error = \frac{(232 - 234) + (255 - 256) + \ldots}{4}$$

$$Error = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

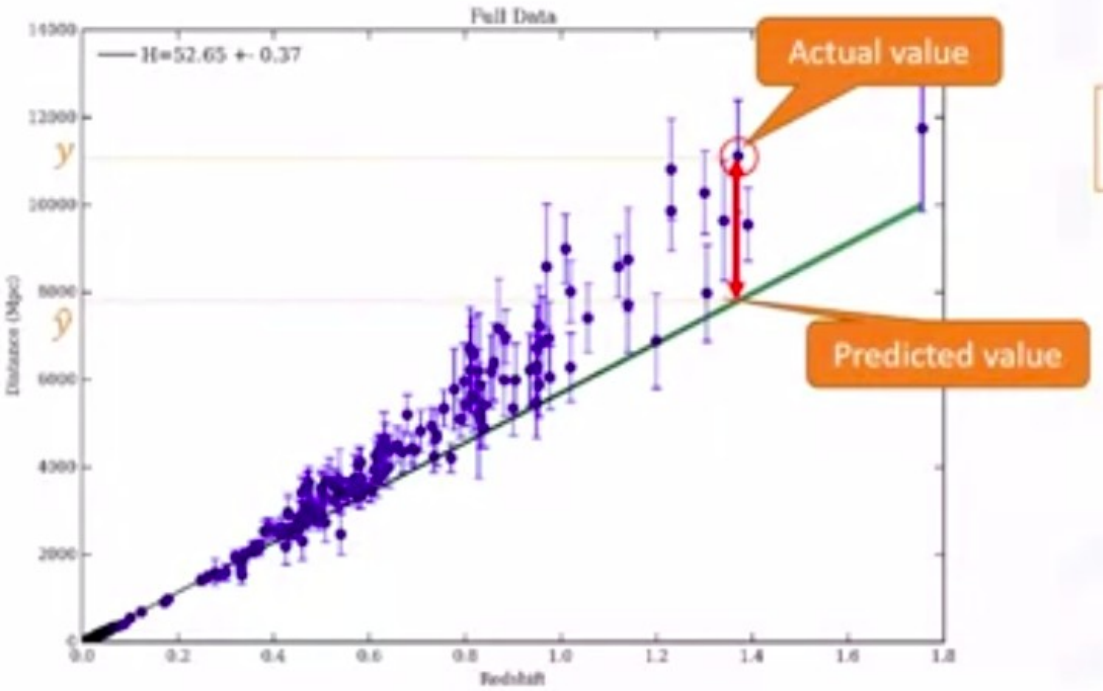
- MAE
- MSE
- RMSE
- ...

$\hat{y}$

	Prediction
6	234
7	256
8	267
9	210

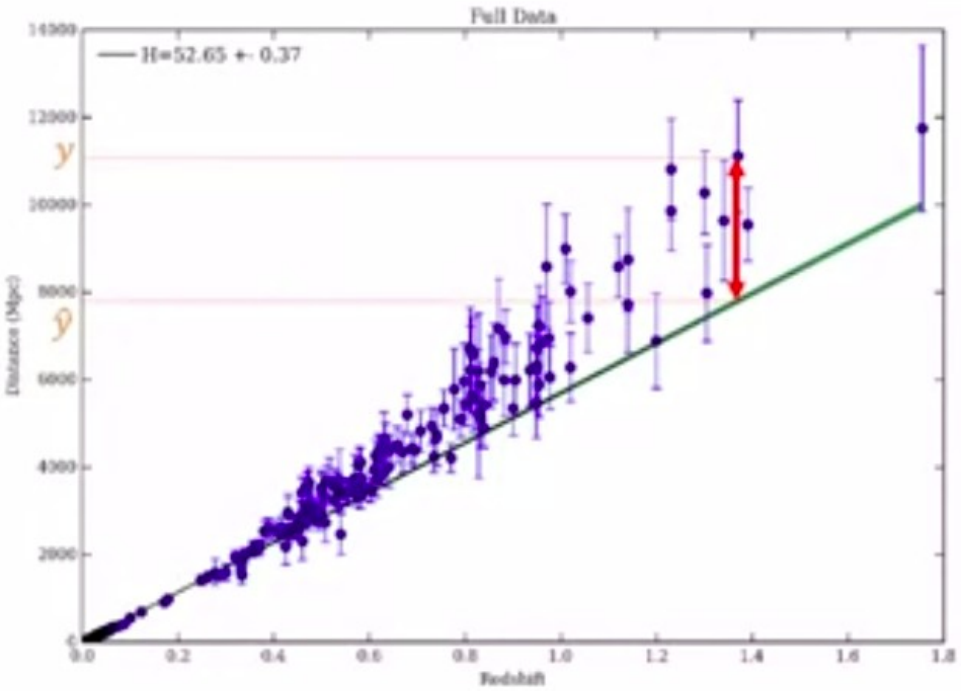
Predicted values

What is an error of the model?



Error: measure of how far the data is from the fitted regression line.

What is an error of the model?



$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

$$RAE = \frac{\sum_{j=1}^{n}|y_j - \hat{y}_j|}{\sum_{j=1}^{n}|y_j - \bar{y}|}$$

$$RSE = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

$$R^2 = 1 - RSE$$

Types of regression models

- Simple Linear Regression

- Multiple Linear Regression

Multiple Linear Regression

Types of regression models

- Simple Linear Regression
  - Predict Co2emission vs EngineSize of all cars
    - Independent variable (x): EngineSize
    - Dependent variable (y): Co2emission
- Multiple Linear Regression
  - Predict Co2emission vs EngineSize and Cylinders of all cars
    - Independent variable (x): EngineSize, Cylinders, etc.
    - Dependent variable (y): Co2emission

Examples of multiple linear regression

- Independent variables effectiveness on prediction
  - Does revision time, test anxiety, lecture attendance and gender have any effect on the exam performance of students?

- Predicting impacts of changes
  - How much does blood pressure go up (or down) for every unit increase (or decrease) in the BMI of a patient?

Predicting continuous values with multiple linear regression

$$Co2\ Em = \theta_0 + \theta_1 Engine\ size + \theta_2 Cylinders + \ ...$$

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ ... + \theta_n x_n$$

$$\hat{y} = \theta^T X$$

$$\theta^T = [\theta_0, \theta_1, \theta_2, ...] \qquad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ ... \end{bmatrix}$$

X: Independent variable

Y: Dependent variable

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Using MSE to expose the errors in the model

$\hat{y} = \theta^T X$

$\hat{y}_i = 140$ the predicted emission of x_i

$y_i = 196$ actual value of x_i

$y_i - \hat{y}_i = 196 - 140 = 56$ residual error

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

Estimating multiple linear regression parameters

- How to estimate θ?
  - Ordinary Least Squares
    - Linear algebra operations
    - Takes a long time for large datasets (10K+ rows)
  - An optimization algorithm
    - Gradient Descent
    - Proper approach if you have a very large dataset

Question

What is the best approach to find the parameter or coefficients for multiple linear regression, when we have very large dataset?

- ○ Using linear algebra operations
- ◉ Using an optimization approach

✓ **Correct**

Skip Continue

Making predictions with multiple linear regression

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

$$\hat{y} = \theta^T X$$

$$\theta^T = [125, 6.2, 14, \dots]$$

$$\hat{y} = 125 + 6.2x_1 + 14x_2 +$$

$$Co2Em = 125 + 6.2EngSize + 14\ \text{Cylinders} + \dots$$

$$Co2Em = 125 + 6.2 \times 2.4 + 14 \times 4 + \dots$$

$$Co2Em = 214.1$$

Back

Practice Quiz: Regression

Practice Quiz • 10 min • 3 total points

1. Which of the following is the meaning of "Out of Sample Accuracy" in the context of evaluation of models?

1 / 1 point

- ○ “Out of Sample Accuracy” is the accuracy of a model on all the data available.
- ◉ "Out of Sample Accuracy" is the percentage of correct predictions that the model makes on data that the model has NOT been trained on.
- ○ "Out of Sample Accuracy" is the accuracy of an overly trained model (which may capture noise and produced a non-generalized model)
- ○ “Out of Sample Accuracy” is the percentage of correct predictions that the model makes using the test dataset.

Correct

Correct! Out-of-sample accuracy represents how well the model is able to perform on unknown data.

2. When should we use Multiple Linear Regression? (Select two)

1 / 1 point

Back

Practice Quiz: Regression

Practice Quiz • 10 min • 3 total points

2. When should we use Multiple Linear Regression? (Select two)

1 / 1 point

- [x] When we would like to predict impacts of changes in independent variables on a dependent variable.

Correct
Correct! We hope to understand how the dependent variable change when we change the independent variables.

- [ ] When we would like to examine the relationship between multiple variables.
- [ ] When there are multiple dependent variables
- [x] When we would like to identify the strength of the effect that the independent variables have on a dependent variable.

Correct
Correct! Multiple linear regression is used for regression tasks involving more than one independent variable.

Correct! Multiple linear regression is used for regression tasks involving more than one independent variable.

3. Which sentence is TRUE about linear regression? **1 / 1 point**

- (●) A linear relationship is necessary between the independent variables and the dependent variable.
- ( ) Multiple linear regression requires a linear relationship between the predictors and the response, but simple linear regression does not.
- ( ) Simple linear regression requires a linear relationship between the predictor and the response, but multiple linear regression does not.
- ( ) A linear relationship is necessary between the independent and dependent variables as well as in between independent variables.

Correct
Correct! If the relationship is non-linear, then we must use non-linear regression.

What is classification?

- A supervised learning approach
- Categorizing some unknown items into a discrete set of categories or "classes"
- The target attribute is a categorical variable

How does classification work?

Classification determines the class label for an unlabeled test case.

age	ed	employ	address	income	debtinc	creddebt	othdebt	default
41	3	17	12	176	9.3	11.359	5.009	1
27	1	10	6	31	17.3	1.362	4.001	0
40	1	15	14	55	5.5	0.856	2.169	0
41	1	15	14	120	2.9	2.659	0.821	0
24	2	2	0	28	17.3	1.787	3.057	1
41	2	5	5	25	10.2	0.393	2.157	0
39	1	20	9	67	30.6	3.834	16.668	0
43	1	12	11	38	3.6	0.129	1.239	0
24	1	3	4	19	24.4	1.358	3.278	1
36	1	0	13	25	19.7	2.778	2.147	0

Categorical Variable

Modeling

age	ed	employ	address	income	debtinc	creddebt	othdebt	default
37	2	16	10	130	9.3	10.23	3.21	0

Prediction

Classifier

Predicted Labels

Question

What is a **multi-class classifier**?

- ○ A classifier that can predict multiple fields with many discrete values.
- ○ A classifier that can predict a field with two discrete values, such as ”Defaulter" or "Not Defaulter”.
- ◉ A classifier that can predict a field with multiple discrete values, such as ”DrugA", ”DrugX" or "DrugY”.

✓ **Correct**

Skip Continue

Example of multi-class classification

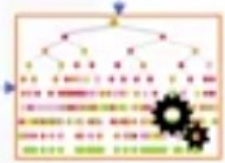
Age	Sex	BP	Cholesterol	Na	K	Drug
23	F	HIGH	HIGH	0.793	0.031	drugY
47	M	LOW	HIGH	0.739	0.056	drugC
47	M	LOW	HIGH	0.697	0.069	drugC
28	F	NORMAL	HIGH	0.564	0.072	drugX
61	F	LOW	HIGH	0.559	0.031	drugY
22	F	NORMAL	HIGH	0.677	0.079	drugX
49	F	NORMAL	HIGH	0.79	0.049	drugY
41	M	LOW	HIGH	0.767	0.069	drugC
60	M	NORMAL	HIGH	0.777	0.051	drugY
43	M	LOW	NORMAL	0.526	0.027	drugY

Categorical Variable

Modeling

Age	Sex	BP	Cholesterol	Na	K	Drug
36	F	LOW	HIGH	0.697	0.069	DrugX

Prediction



Classifier

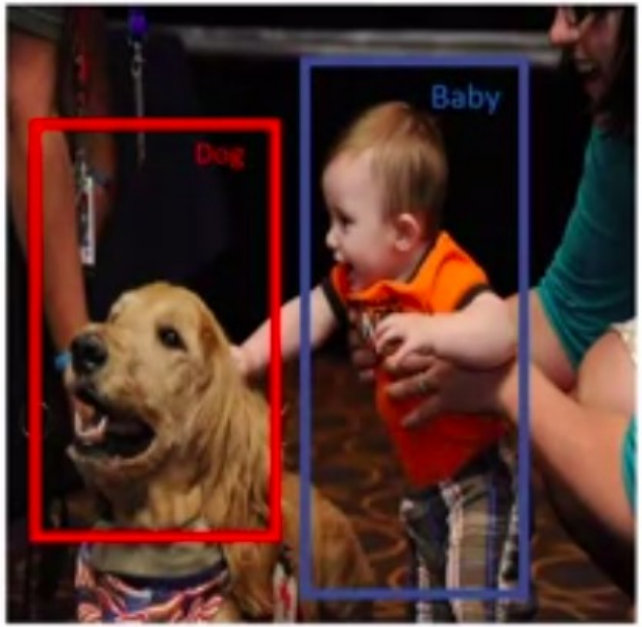
Predicted Labels

Classification use cases

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	?

- Which category a customer belongs to?
- Whether a customer switches to another provider/brand?
- Whether a customer responds to a particular advertising campaign?

Classification applications



Classification algorithms in machine learning

- Decision Trees (ID3, C4.5, C5.0)
- Naïve Bayes
- Linear Discriminant Analysis
- *k*-Nearest Neighbor
- Logistic Regression
- Neural Networks
- Support Vector Machines (SVM) 


Press Esc to exit full screen
K-Nearest Neighbours
IBM Developer
SKILLS NETWORK
0:06 / 9:17

Intro to KNN

X: Independent variable

Y: Dependent variable

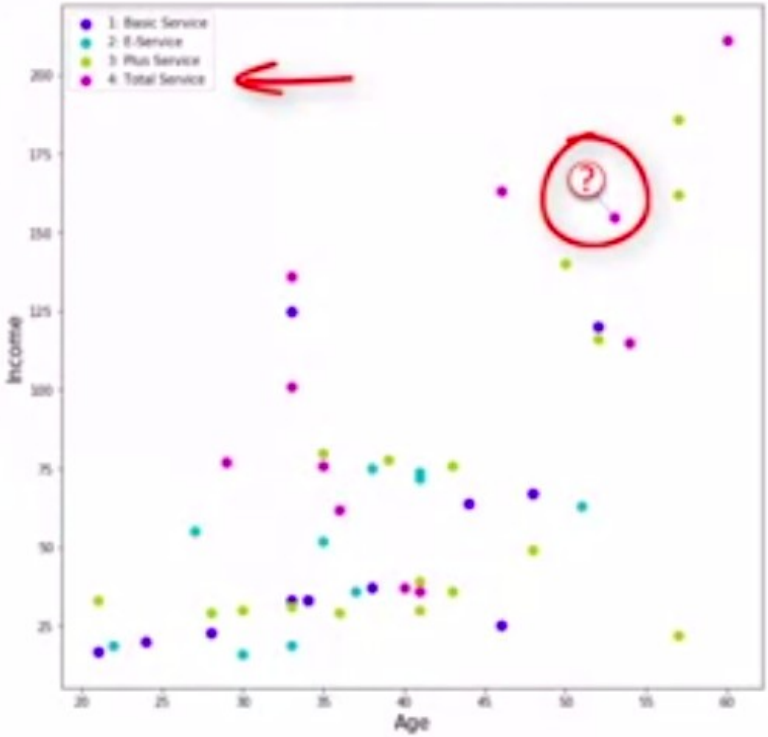
	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

Value	Label
1	Basic Service
2	E-Service
3	Plus Service
4	Total Service

Determining the class using 1st KNN

	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	70	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

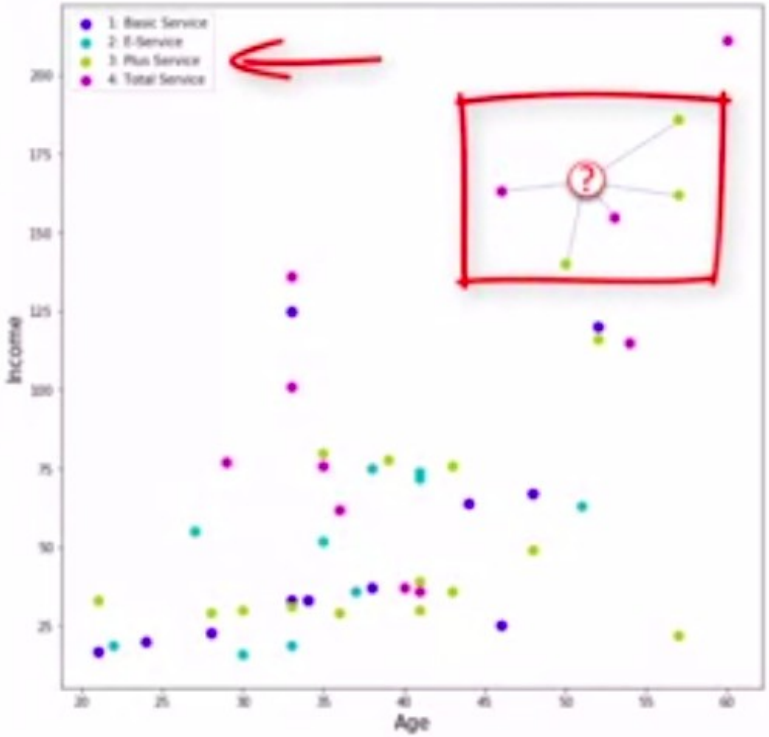
1-NN → 4: Total Service



Determining the class using the 5 KNNs

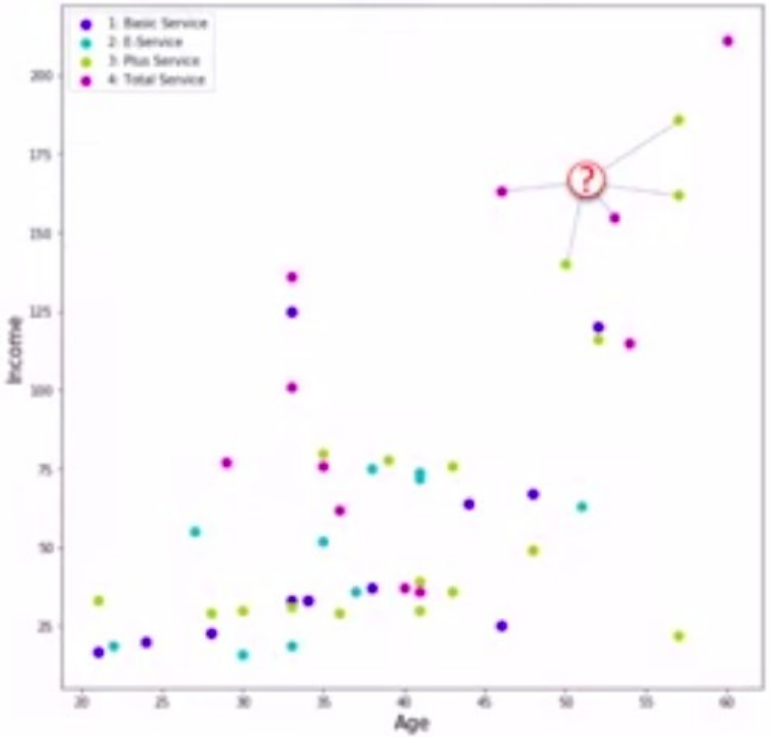
	region	age	marital	address	income	ed	employ	retire	gender	reside	custcat
0	2	44	1	9	64	4	5	0	0	2	1
1	3	33	1	7	136	5	5	0	0	6	4
2	3	52	1	24	116	1	29	0	1	2	3
3	2	33	0	12	33	2	0	0	1	1	1
4	2	30	1	9	30	1	2	0	0	4	3
5	2	39	0	17	78	2	16	0	1	1	3
6	3	22	1	2	19	2	4	0	1	5	2
7	2	35	0	5	76	2	10	0	0	3	4
8	3	50	1	7	166	4	31	0	0	5	?

5-NN → 3: Plus Service



What is K-Nearest Neighbor (or KNN)?

- A method for classifying cases based on their similarity to other cases
- Cases that are near each other are said to be "neighbors"
- Based on similar cases with same class labels are near each other



The K-Nearest Neighbors algorithm

1. Pick a value for K.
2. Calculate the distance of unknown case from all cases.
3. Select the K-observations in the training data that are “nearest” to the unknown data point.
4. Predict the response of the unknown data point using the most popular response value from the K-nearest neighbors.

Calculating the similarity/distance in a 1-dimensional space

Customer 1
Age
34

Customer 2
Age
30

$$\text{Dis}\,(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

$$\text{Dis}\,(x_1, x_2) = \sqrt{(34 - 30)^2} = 4$$

Calculating the similarity/distance in a multi-dimensional space

Customer 1		
Age	Income	Education
34	190	3

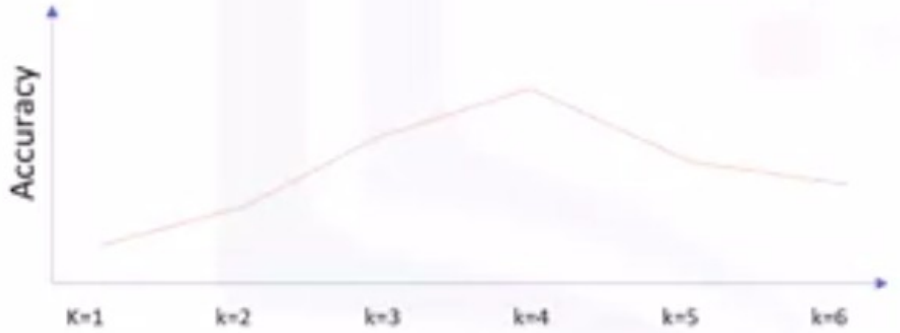
Customer 2		
Age	Income	Education
30	200	8

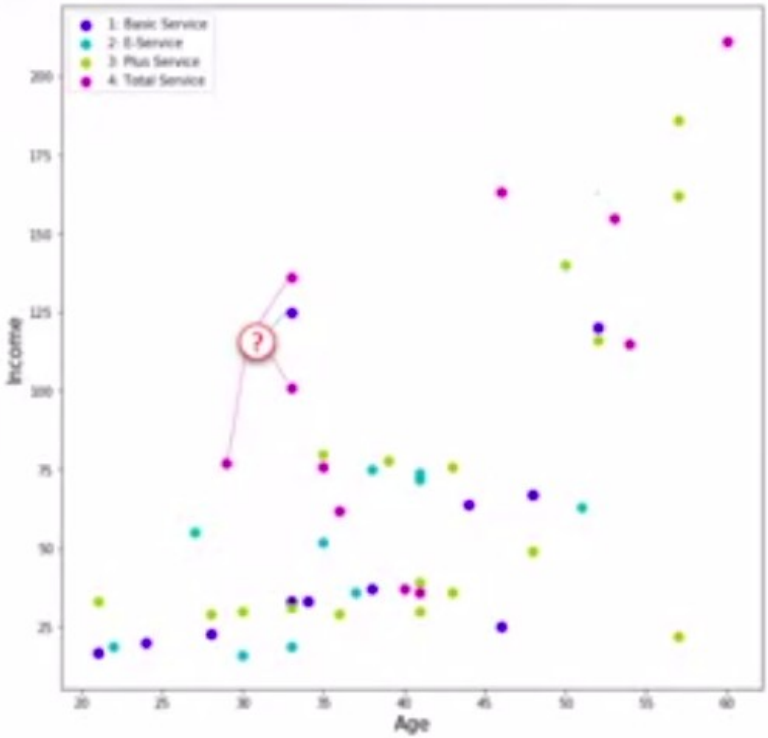
$$\text{Dis}\,(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

$$= \sqrt{(34-30)^2+(190-200)^2+(3-8)^2} = 11.87$$

What is the best value of K for KNN?

- K =1 class 1
- K =20 ?





Question

Which of the following statements are TRUE about kNN?

- [x] The kNN algorithm is a classification algorithm.

Correct

- [x] The kNN algorithm classify cases based on their similarity to other cases.

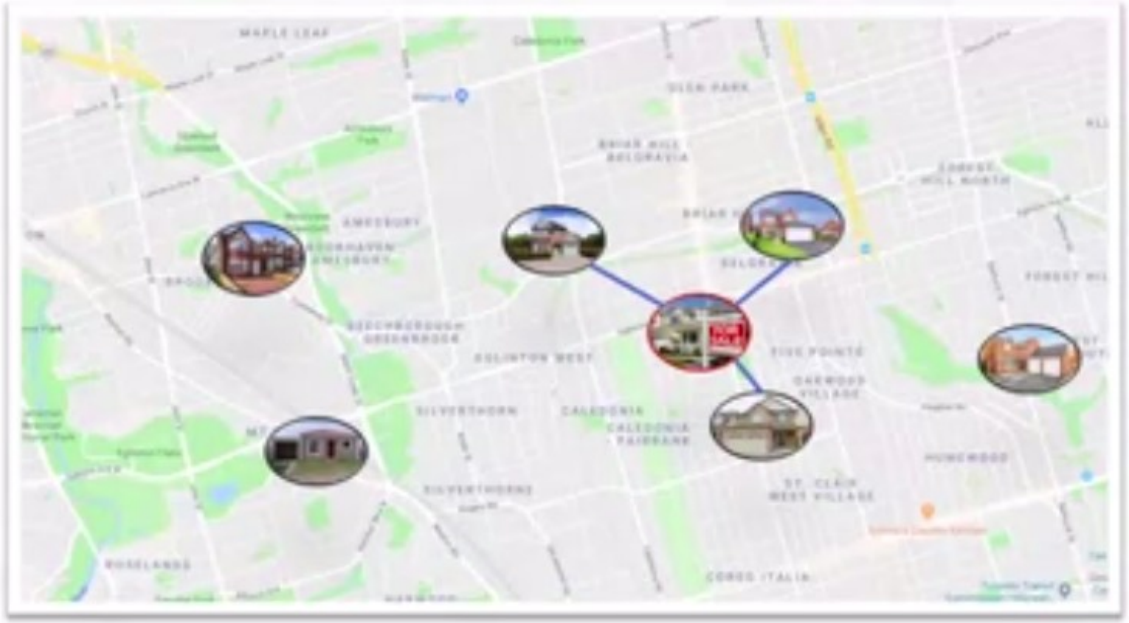
Correct

- [ ] The kNN algorithm works only with Euclidian distance.
- [ ] The bigger K is in KNN, the accuracy of the algorithm is better.

Skip Continue

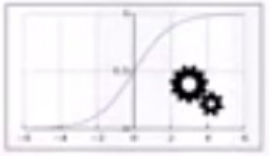
Computing continuous targets using KNN

- KNN can also be used for regression



Evaluation Metrics in Classification

Classification accuracy



Test set

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0

Test

Predicted churn
1
1
0
1
0

Actual Labels y

$\hat{y}$ Predicted Labels

Jaccard index

y: Actual labels

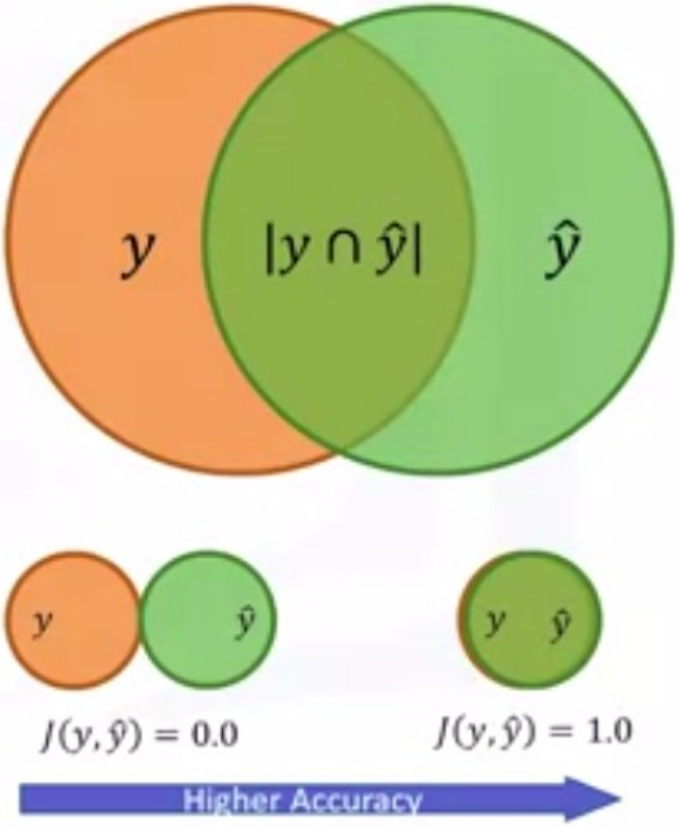
$\hat{y}$: Predicted labels

$$J(y,\hat{y}) = \frac{|y \cap \hat{y}|}{|y \cup \hat{y}|} = \frac{|y \cap \hat{y}|}{|y| + |\hat{y}| - |y \cap \hat{y}|}$$

y: [0, 0, 0, 0, 0, 1, 1, 1, ,1 ,1]

$\hat{y}$: [1, 1, 0, 0, 0, 1, 1, 1, ,1 ,1]

$$J(y,\hat{y}) = \frac{8}{10+10-8} = 0.66$$



F1-score

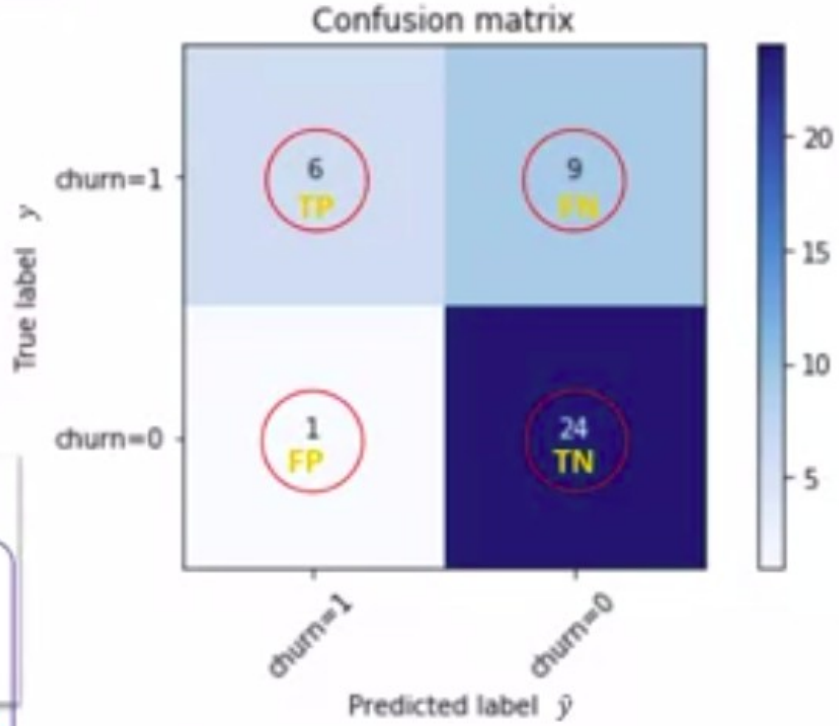
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1-score = 2x (prc x rec) / (prc+rec)

F1-score: 0.00 ... 0.20 .... 0.55 .... 0.83 ... 1.00

Higher Accuracy

	precision	recall	f1-score
Churn = 0	0.73	0.96	0.83
Churn = 1	0.86	0.40	0.55

Avg Accuracy = 0.72



Question

Which one is the ideal classifier?

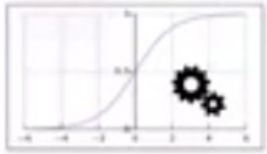
- (●) The classifier with F1-score close to one.
- ( ) The classifier with LogLoss close to one.
- ( ) The classifier with Jaccard-index close to zero.

✓ **Correct**

Skip | Continue

Log loss

Performance of a classifier where the predicted output is a probability value between 0 and 1.



Test set

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0

Test

Predicted churn	LogLoss
0.91	0.094
0.13	0.89
0.04	0.04
0.23	0.26
0.43	0.56

$LogLoss = 0.60$

Actual Labels y

$\hat{y}$ Predicted Probability

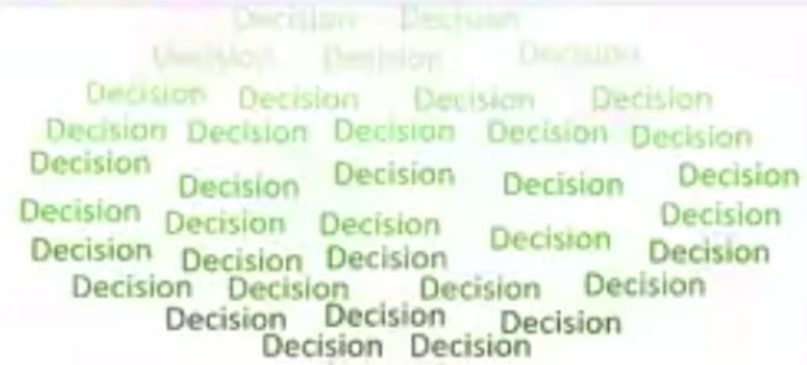
$$LogLoss = -\frac{1}{n}\sum (y \times \log(\hat{y}) + (1-y) \times \log(1-\hat{y}))$$

LogLoss: 0.00 ... 0.35 .... 0.60 ... 1.00

Higher Accuracy

Introduction to Decision Trees

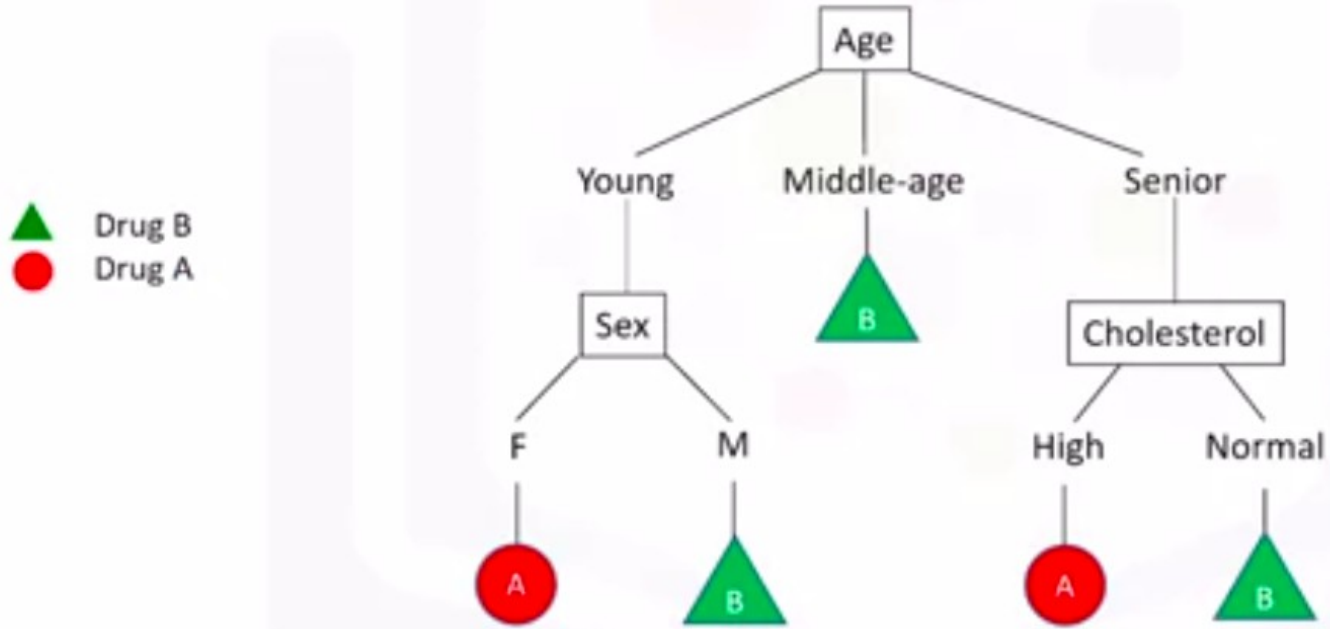
What is a decision tree?



The basic intuition behind a decision tree is to map out all possible decision paths in the form of a tree.

Narendra Nath Joshi

Building a decision tree with the training set



How to build a decision tree?

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A
p15	Middle-age	F	Low	Normal	?

Question

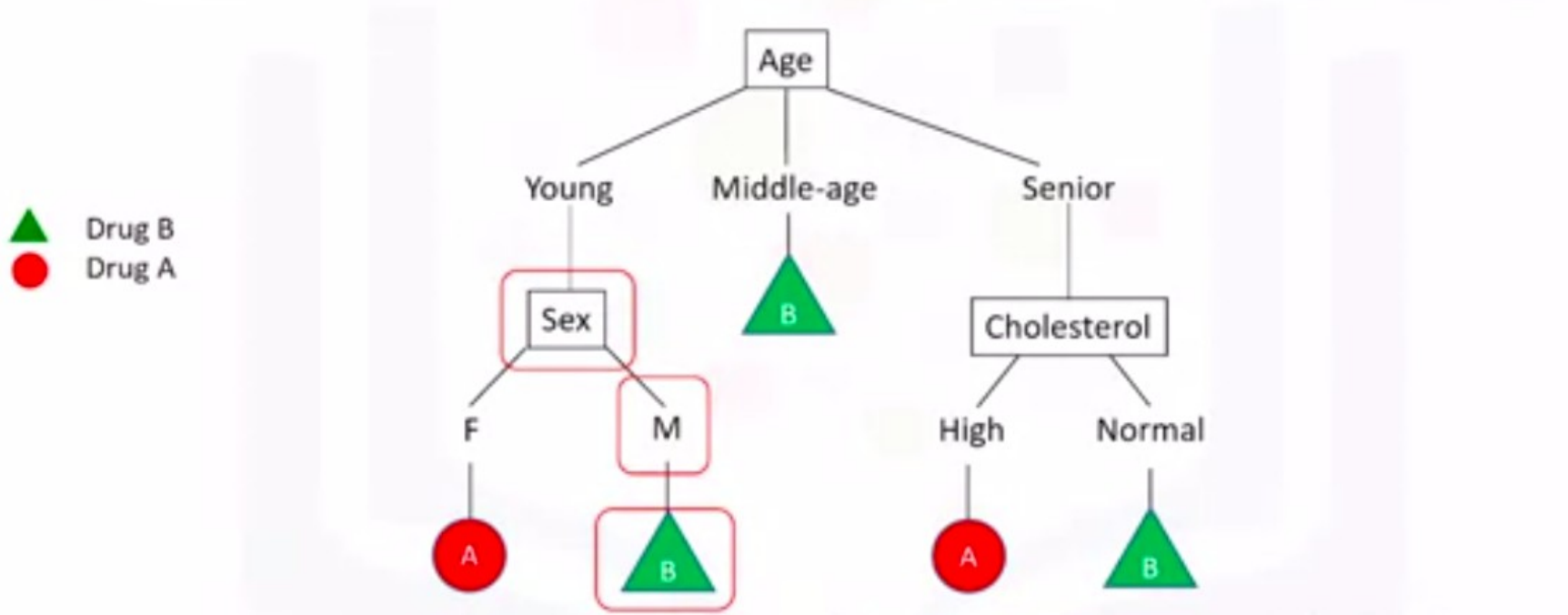
Which of the following sentences is **NOT TRUE** about Decision Tree?

- ( ) Decision Trees are built by splitting the training set into distinct nodes
- (•) A Decision Tree is a type of clustering approach that can predict the class of a group, for example, DrugA or DrugB.
- ( ) One node in a Decision Tree contains all of or most of, one category of the data.

Correct

Skip Continue

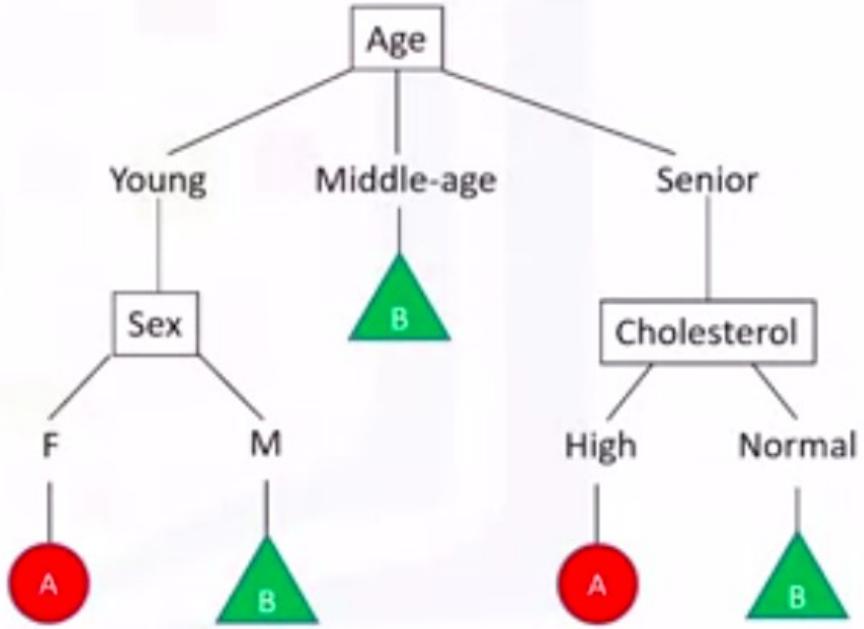
Building a decision tree with the training set



- Each internal node corresponds to a test
- Each branch corresponds to a result of the test
- Each leaf node assigns a classification

Decision tree learning algorithm

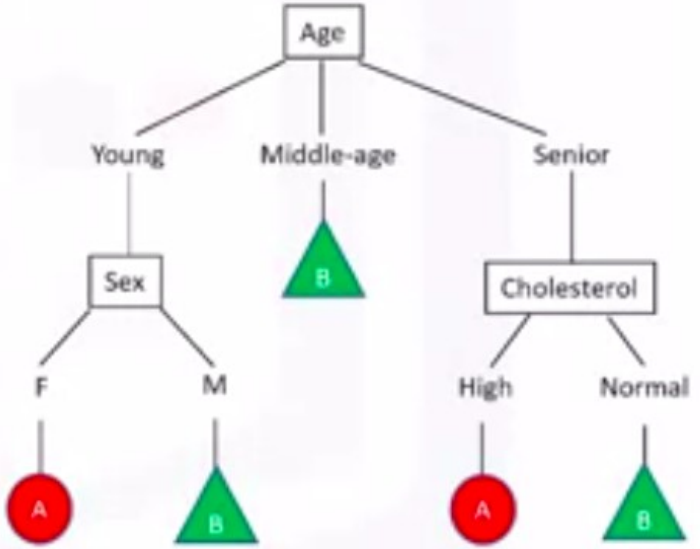
1. Choose an attribute from your dataset.
2. Calculate the significance of attribute in splitting of data.
3. Split data based on the value of the best attribute.
4. Go to step 1.



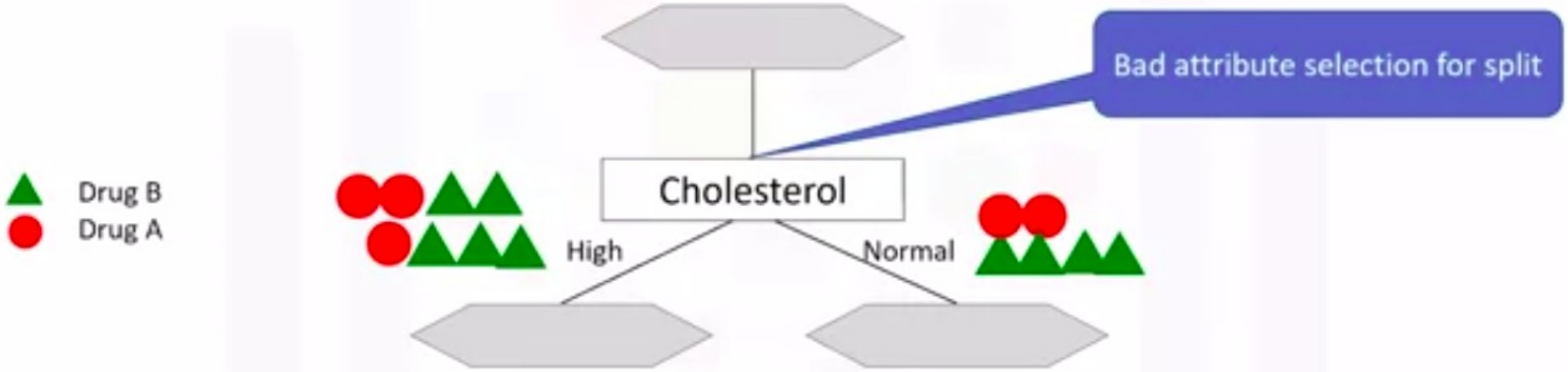
Building Decision Trees

How do you build a decision tree?

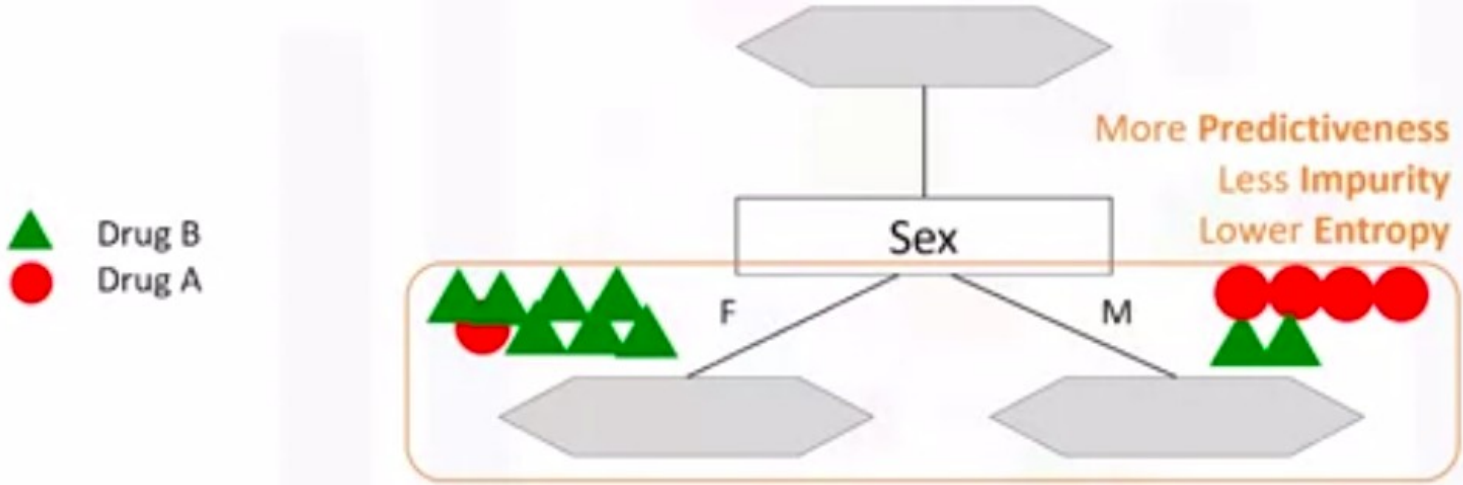
Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A
p15	Middle-age	F	Low	Normal	?



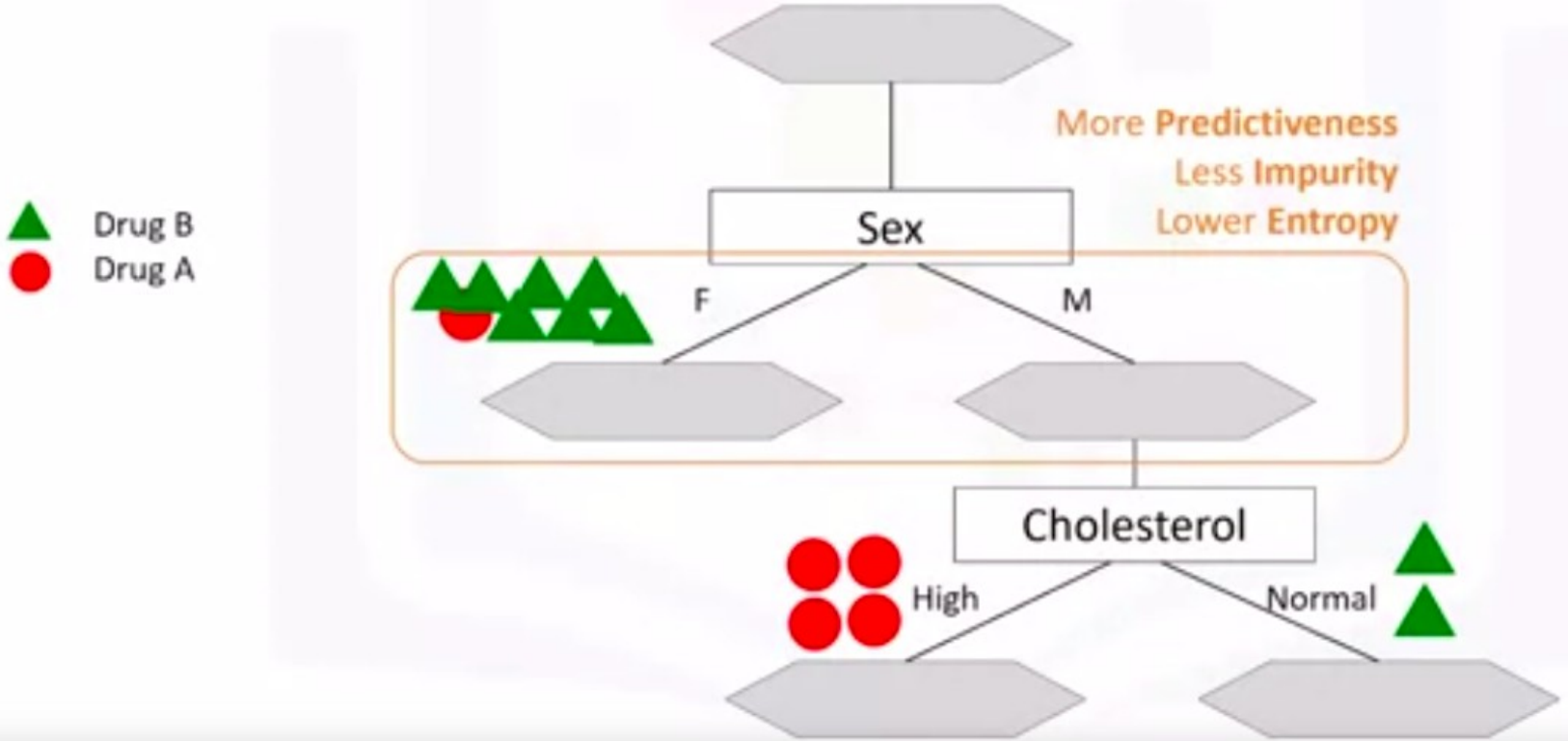
Which attribute is the best ?



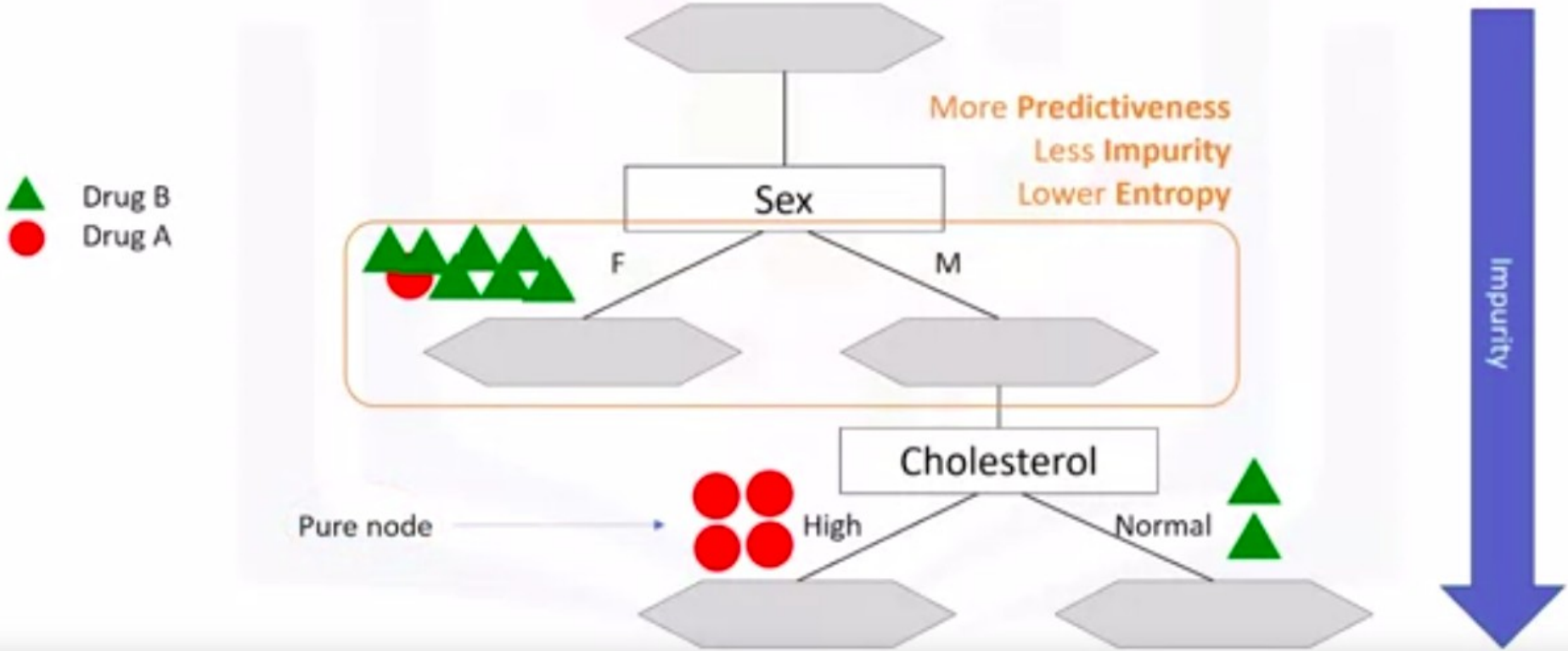
Which attribute is the best ?



Which attribute is the best ?



Which attribute is the best ?

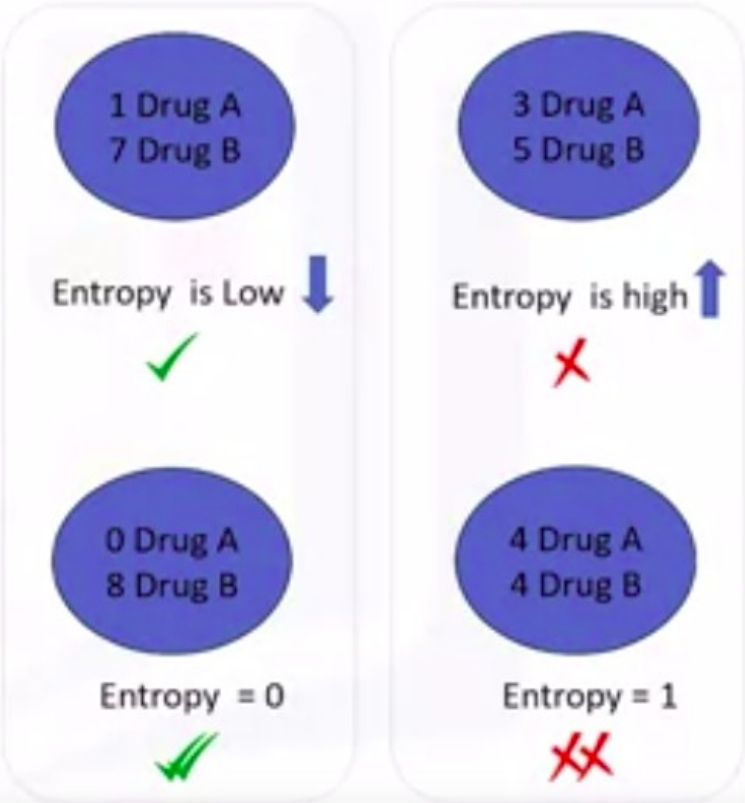


Entropy

- Measure of randomness or uncertainty

$$\text{Entropy} = - p(A)\log_2(p(A)) - p(B)\log_2(p(B))$$

The lower the Entropy, the less uniform the distribution, the purer the node.



Which attribute is the best one to use?

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

S: [9 B, 5 A]

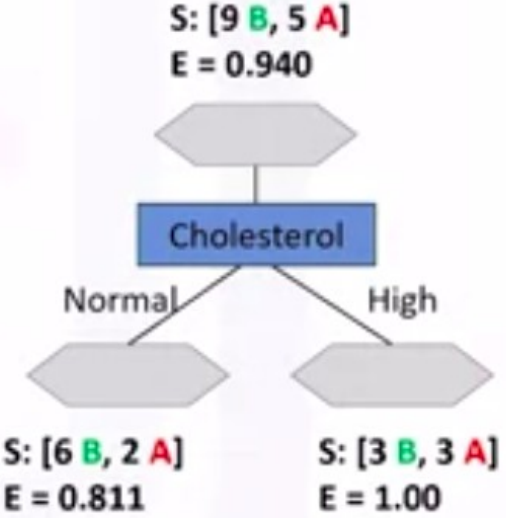
$E = - p(B)\log_2(p(B)) - p(A)\log_2(p(A))$

$E = - (9/14)\log_2(9/14) - (5/14)\log_2(5/14$

E = 0.940

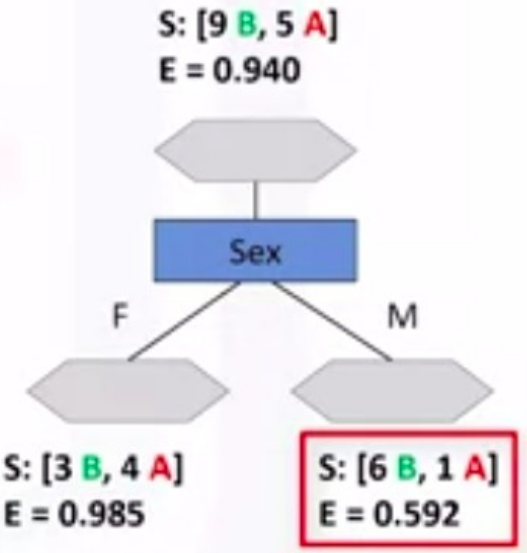
Is 'Cholesterol' the best attribute?

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

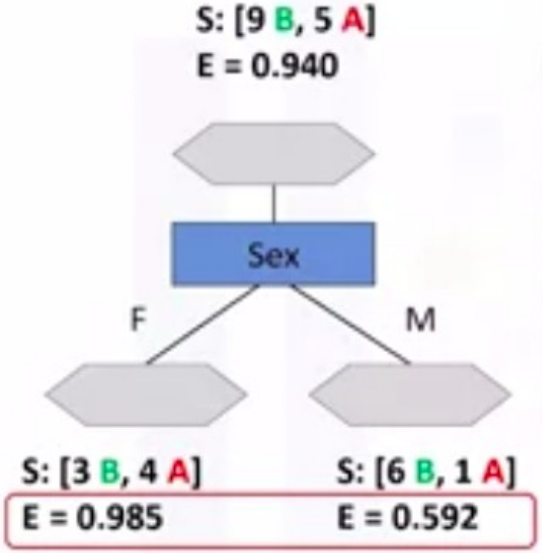


What about 'Sex'?

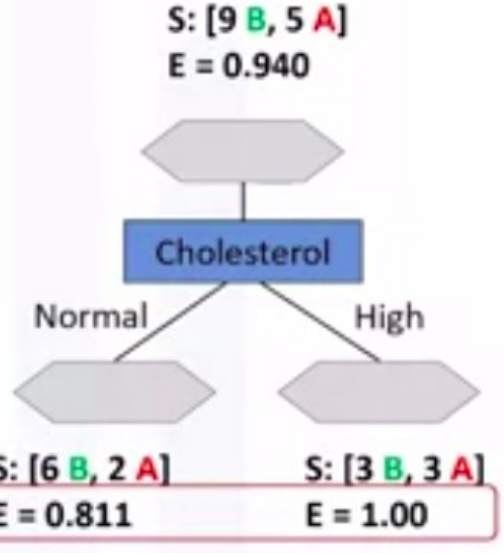
Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	Hiigh	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A



Which attribute is the best?



Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	High	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A



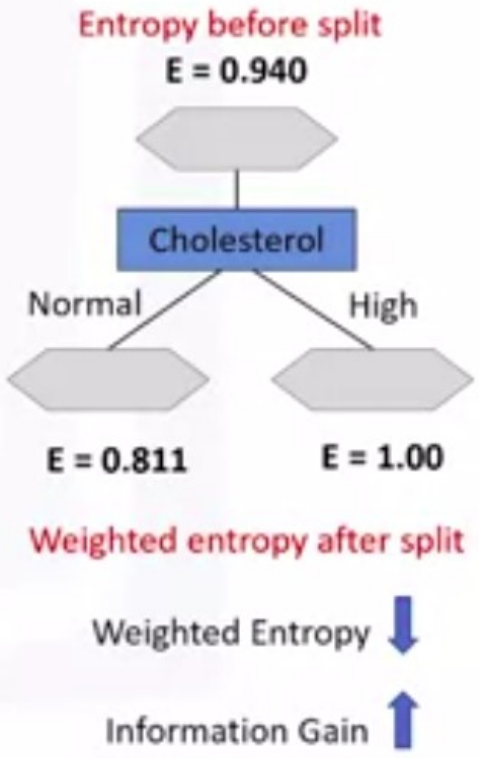
?

The tree with the higher Information Gain after splitting.

What is information gain?

Information gain is the information that can increase the level of certainty after splitting.

Information Gain = (Entropy before split) – (weighted entropy after split)



Question

What is the meaning of **Entropy** in Decision Tree?

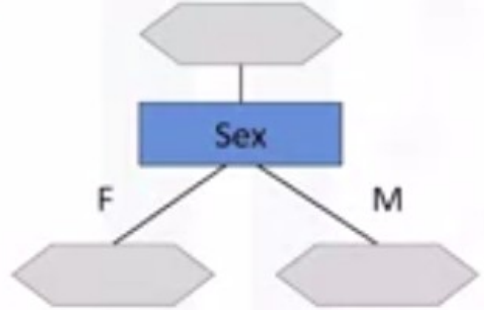
- (●) The entropy in a node is the amount of information disorder calculated in each node.
- ( ) The entropy in a node is the number of similar data in that node.
- ( ) The entropy in a node is the weighted information in its parent node.

✓ **Correct**

Skip Continue

Which attribute is the best?

S: [9 B, 5 A]
E = 0.940



S: [3 B, 4 A]
E = 0.985

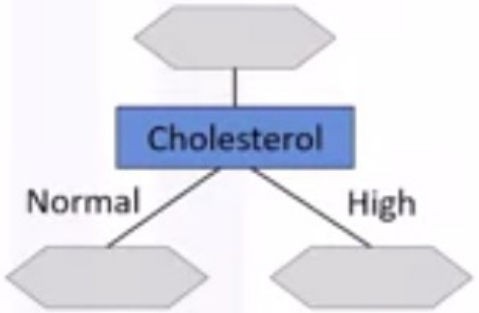
S: [6 B, 1 A]
E = 0.592

Gain (s, Sex)
= 0.940 – [(7/14)0.985 + (7/14)0.592]
=0.151

Patient ID	Age	Sex	BP	Cholesterol	Drug
p1	Young	F	High	Normal	Drug A
p2	Young	F	High	High	Drug A
p3	Middle-age	F	High	Normal	Drug B
p4	Senior	F	Normal	Normal	Drug B
p5	Senior	M	Low	Normal	Drug B
p6	Senior	M	Low	High	Drug A
p7	Middle-age	M	Low	High	Drug B
p8	Young	F	Normal	Normal	Drug A
p9	Young	M	Low	Normal	Drug B
p10	Senior	M	Normal	Normal	Drug B
p11	Young	M	Normal	High	Drug B
p12	Middle-age	F	Normal	High	Drug B
p13	Middle-age	M	High	Normal	Drug B
p14	Senior	F	Normal	High	Drug A

?

S: [9 B, 5 A]
E = 0.940

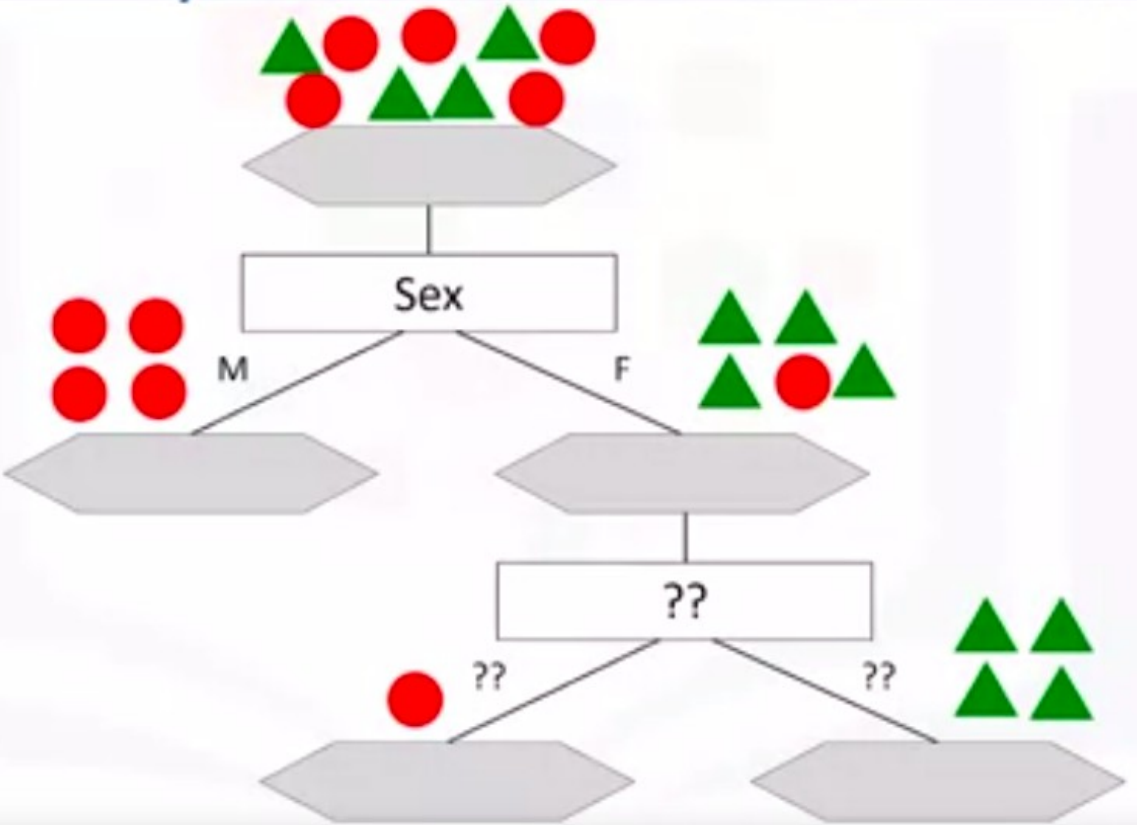


S: [6 B, 2 A]
E = 0.811

S: [3 B, 3 A]
E = 1.00

Gain (s, Cholesterol)
= 0.940 – [(8/14).811 + (6/14)1.0]
=0.048

Correct way to build a decision tree



Back

Practice Quiz: Classification

Practice Quiz • 10 min • 3 total points

1. Which one is TRUE about the kNN algorithm? **1 / 1 point**

- (●) kNN algorithm can be used to estimate values for a continuous target.
- ( ) kNN calculates similarity by measuring how close the two data points' response values are.
- ( ) kNN is a classification algorithm that takes a bunch of unlabelled points and uses them to learn how to label other points.
- ( ) The most similar point in kNN is the one with the smallest distance averaged across all normalized features.

Correct

Correct! kNN can be used for both classification and regression prediction tasks. In the case of a continuous target, the prediction is taken as the average or median of the nearest neighbours.

2. If the information gain of the tree by using attribute A is 0.3, what can we infer? **1 / 1 point**

- Compared to attribute B with 0.65 information gain, attribute A should be selected first for splitting.
- By making this split, we increase the randomness in each child node by 0.3.
- (selected) The entropy of a tree before split minus weighted entropy after split by attribute A is 0.3.
- Entropy in the decision tree increases by 0.3 if we make this split.

Correct
Correct! This describes how information gain is calculated, measuring how much certainty has increased by making a split.

3. When we have a value of K for KNN that's too small, what will the model most likely look like? **1 / 1 point**

Correct
Correct! This describes how information gain is calculated, measuring how much certainty has increased by making a split.

3. When we have a value of K for KNN that's too small, what will the model most likely look like? **1 / 1 point**

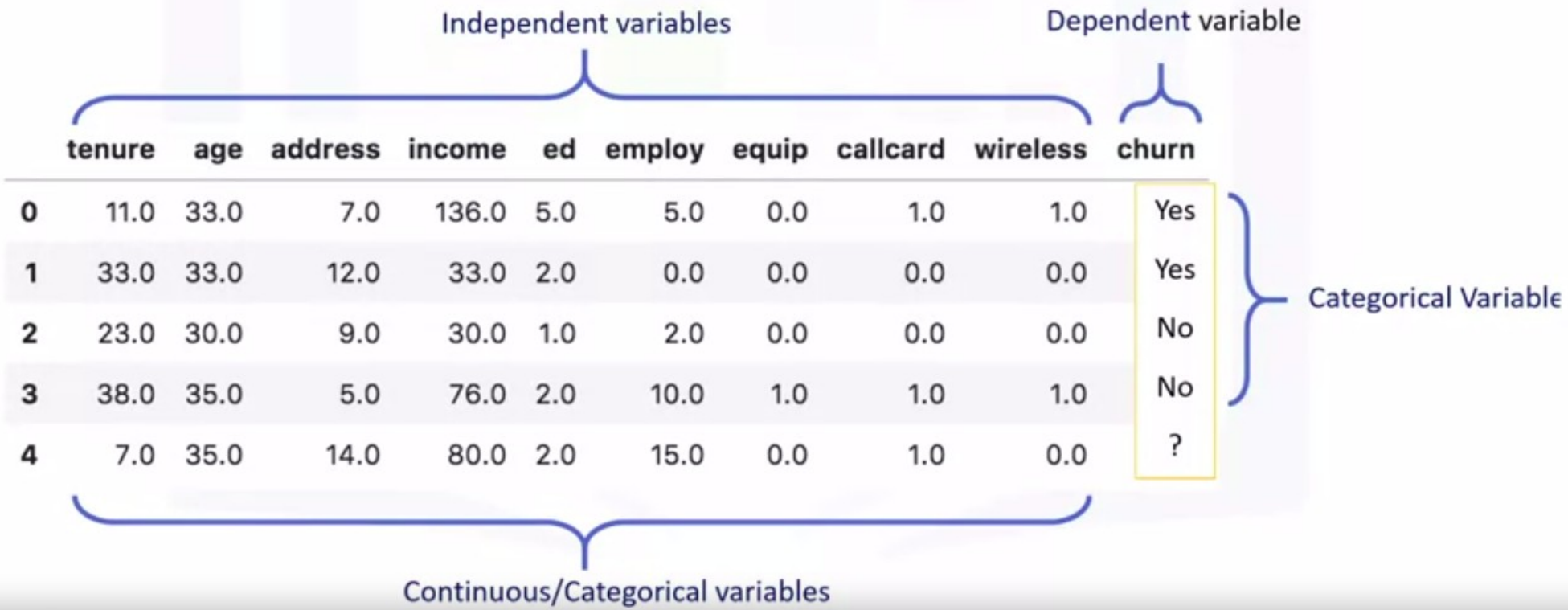
- The model will be overly simple and does not capture enough noise.
- The model will have high out-of-sample accuracy.
- The model will have high accuracy on the test set.
- (selected) The model will be highly complex and captures too much noise.

Correct
Correct! By looking at too few neighbours, we can capture an anomaly in the data, which means that prediction isn't generalized enough.

Intro to Logistic Regression

What is logistic regression?

Logistic regression is a classification algorithm for categorical variables.



	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	?

Logistic regression applications

- Predicting the probability of a person having a heart attack
- Predicting the mortality in injured patients
- Predicting a customer's propensity to purchase a product or halt a subscription
- Predicting the probability of failure of a given process or product
- Predicting the likelihood of a homeowner defaulting on a mortgage

Question

Which of the following sentences are **TRUE** about **Logistic Regression**?

- [x] Logistic regression is analogous to linear regression but takes a categorical/discrete target field instead of a numeric one.

Correct

- [x] Logistic Regression measures the probability of a case belonging to a specific class.

Correct

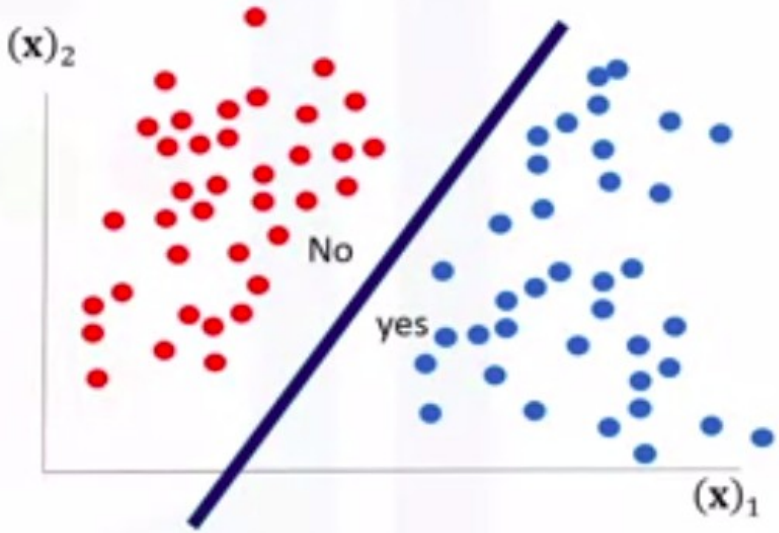
- [x] Logistic Regression can be used to understand the impact of a feature on a dependent variable.

Correct

Skip Continue

When is logistic regression suitable?

- If your data is binary
  - 0/1, YES/NO, True/False
- If you need probabilistic results
- When you need a linear decision boundary
- If you need to understand the impact of a feature



Building a model for customer churn

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No

Building a model for customer churn

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1.0
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1.0
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0.0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0.0

$X \in \mathbb{R}^{m \times n}$

$y \in \{0,1\}$

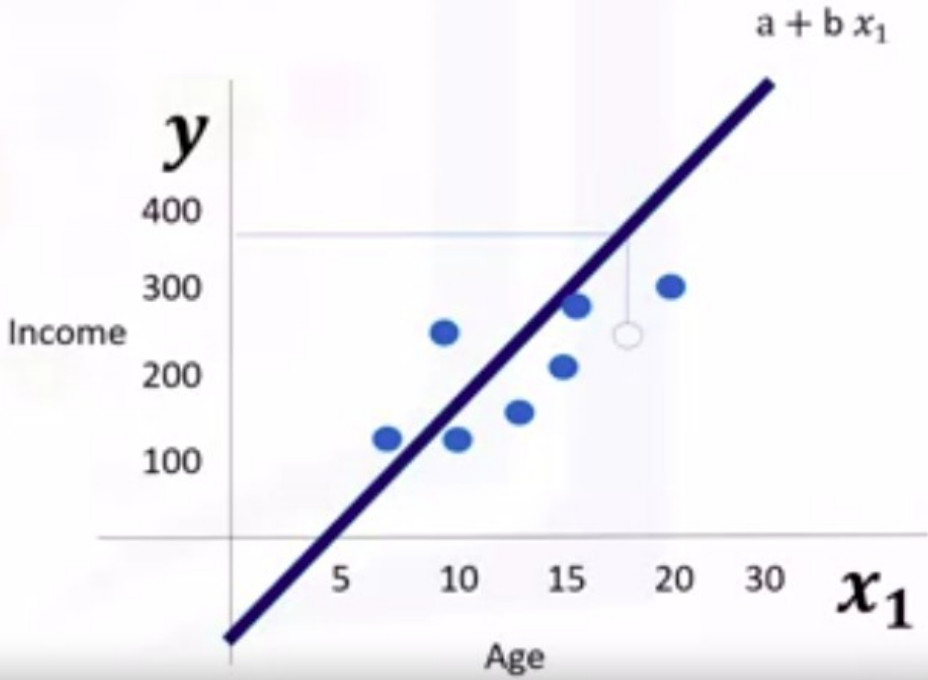
$\hat{y} = \text{P(y=1|x)}$

P(y=0|x) = 1- P(y=1|x)

Logistic Regression vs Linear Regression

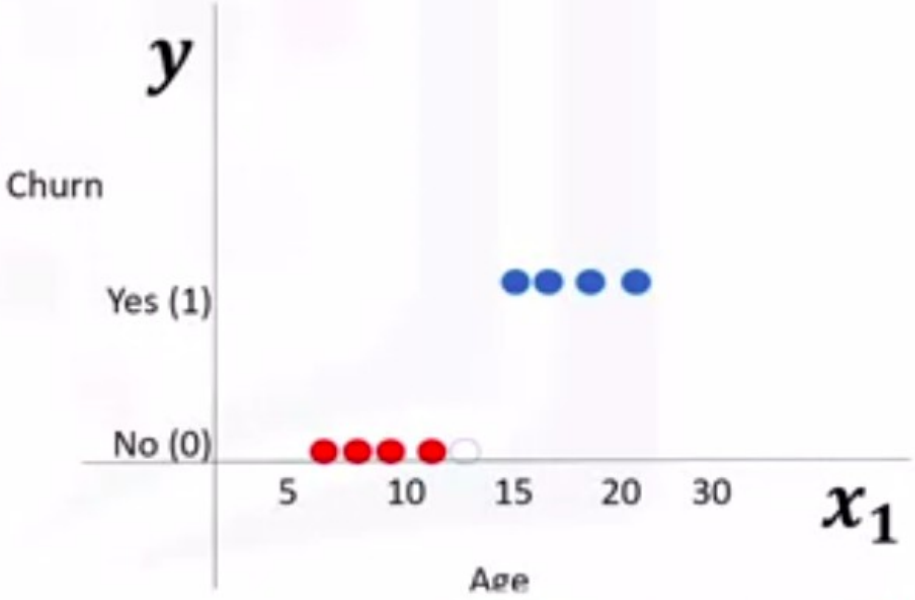
Predicting customer income

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0



Predicting churn using linear regression

	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	1
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	1
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	0
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	0
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	0

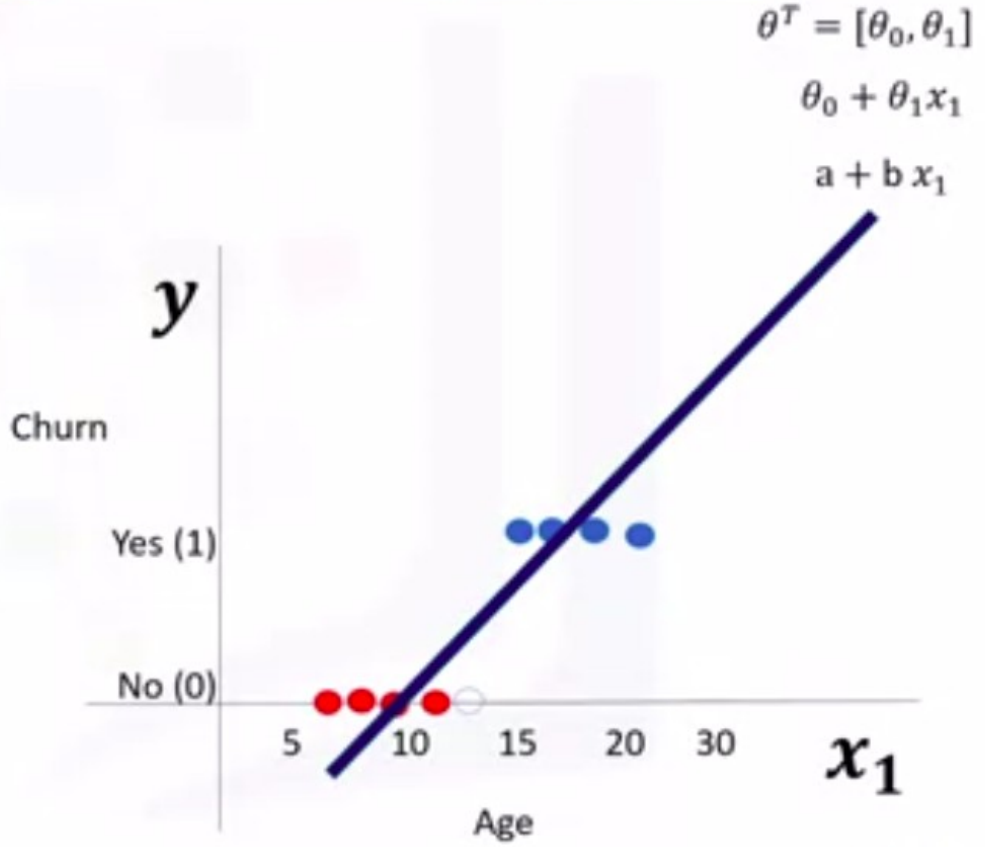


Predicting churn using linear regression

$$\theta^T X = \theta_0 + \theta_1 x_1$$

$$\theta^T X = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots$$

$$\theta^T = [\theta_0, \theta_1, \theta_2, \ldots] \qquad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \ldots \end{bmatrix}$$

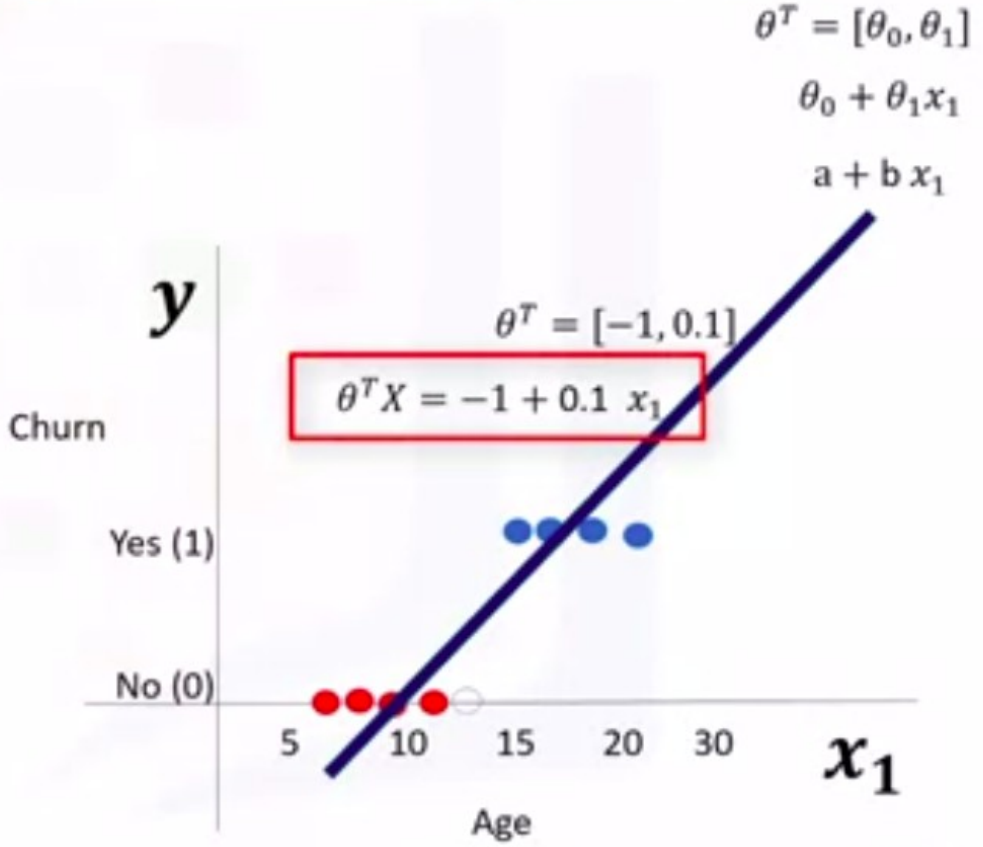


Predicting churn using linear regression

$$\theta^T X = \theta_0 + \theta_1 x_1$$

$$\theta^T X = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots$$

$$\theta^T = [\theta_0, \theta_1, \theta_2, \ldots] \qquad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \ldots \end{bmatrix}$$



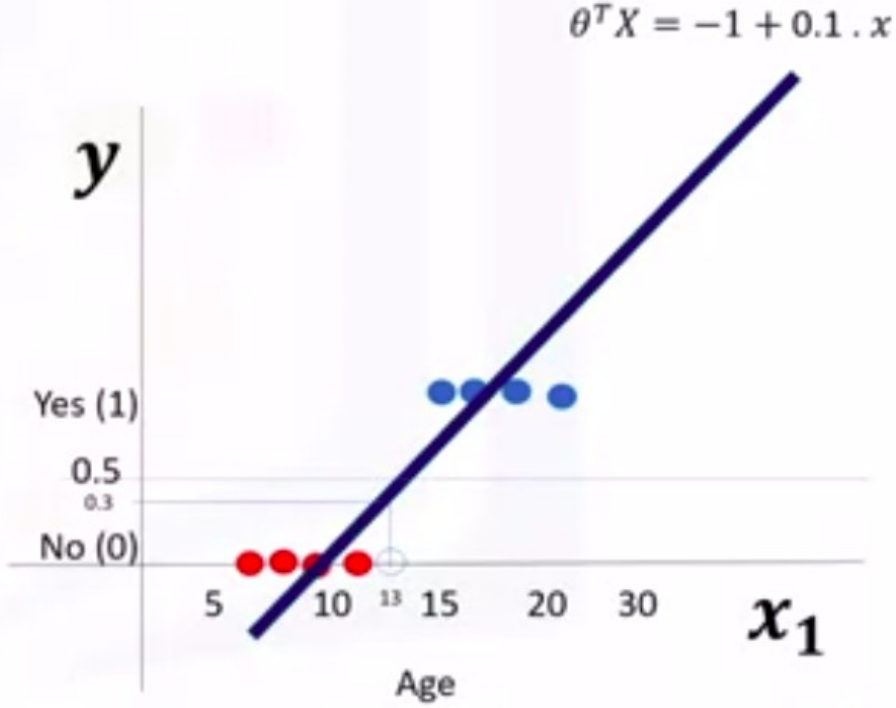
Linear regression in classification problems?

$\theta^T X = \theta_0 + \theta_1 x_1$

$p_1 = [13]$ → $\theta^T X = -1 + 0.1 \,.\, x_1$
= -1 + 0.1 x 13
= 0.3

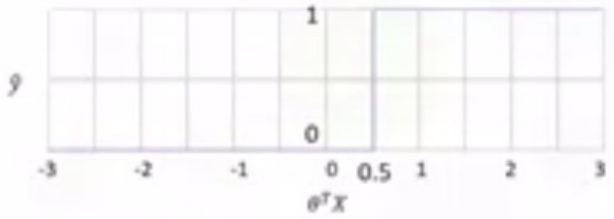
$$\hat{y} = \begin{cases} 0 & \text{if } \theta^T X < 0.5 \\ 1 & \text{if } \theta^T X \geq 0.5 \end{cases}$$

$\theta^T X = 0.3$
$\theta^T X < 0.5$ → Class 0



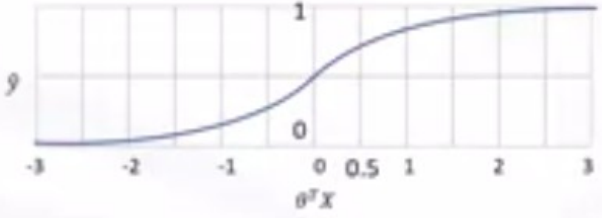
The problem with using linear regression

$$\theta^T X = \theta_0 + \theta_1 x_1 + \cdots$$



$$\hat{y} = \begin{cases} 0 & \text{if } \theta^T X < 0.5 \\ 1 & \text{if } \theta^T X \geq 0.5 \end{cases}$$

$$\sigma(\theta^T X) = \sigma(\theta_0 + \theta_1 x_1 + \cdots)$$



$$\hat{y} = \sigma(\theta^T X)$$

P(y=1|x)

Sigmoid function in logistic regression

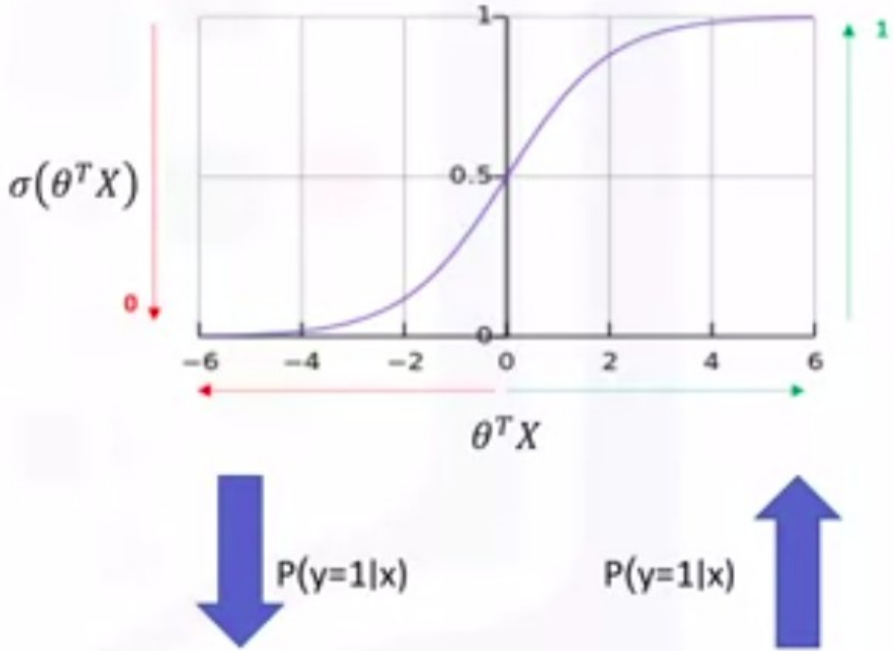
- Logistic Function

$$\sigma(\theta^T X) = \frac{1}{1 + e^{-\theta^T X}}$$

$\sigma(\theta^T X) = 1$

$\sigma(\theta^T X) = 0$

$[0, 1]$



Clarification of the customer churn model

What is the output of our model?

- P(Y=1|X)
- P(y=0|X) = 1 - P(y=1|x)

- P(Churn=1|income,age) = 0.8
- P(Churn=0|income,age) = 1 - 0.8 = 0.2

$\sigma(\theta^T X)$ → P(y=1|x)

$1 - \sigma(\theta^T X)$ → P(y=0|x)

Clarification of the customer churn model

What is the output of our model?

- P(Y=1|X)
- P(y=0|X) = 1 - P(y=1|x)

- P(Churn=1|income,age) = 0.8
- P(Churn=0|income,age) = 1 - 0.8 = 0.2

$\sigma(\theta^T X)$ → P(y=1|x)

$1 - \sigma(\theta^T X)$ → P(y=0|x)

Question

What is difference between Linear Regression vs Logistic Regression, in solving a classification problem?

- (●) Linear Regression cannot properly measure the probability of a case belonging to a class.
- ( ) Linear Regression is very slow in estimating the parameters of the model
- ( ) Linear Regression cannot handle large datasets.

✓ **Correct**

Skip Continue

The training process

$\sigma(\theta^T X)$ → P(y=1|x)

1. Initialize θ.
2. Calculate $\hat{y} = \sigma(\theta^T X)$ for a customer.
3. Compare the output of $\hat{y}$ with actual output of customer, y, and record it as error.
4. Calculate the error for all customers.
5. Change the θ to reduce the cost.
6. Go back to step 2.

$\theta = [-1,2]$

$\hat{y} = \sigma([-1,2] \times [2,5]) = 0.7$

Error = 1-0.7 = 0.3

$Cost = J(\theta)$

θ_{new}

Logistic Regression Training

General cost function

$\sigma(\theta^T X)$ ⟶ P(y=1|x)

- Change the weight -> Reduce the cost

- Cost function

$$Cost(\hat{y}, y) = \frac{1}{2}(\sigma(\theta^T X) - y)^2$$

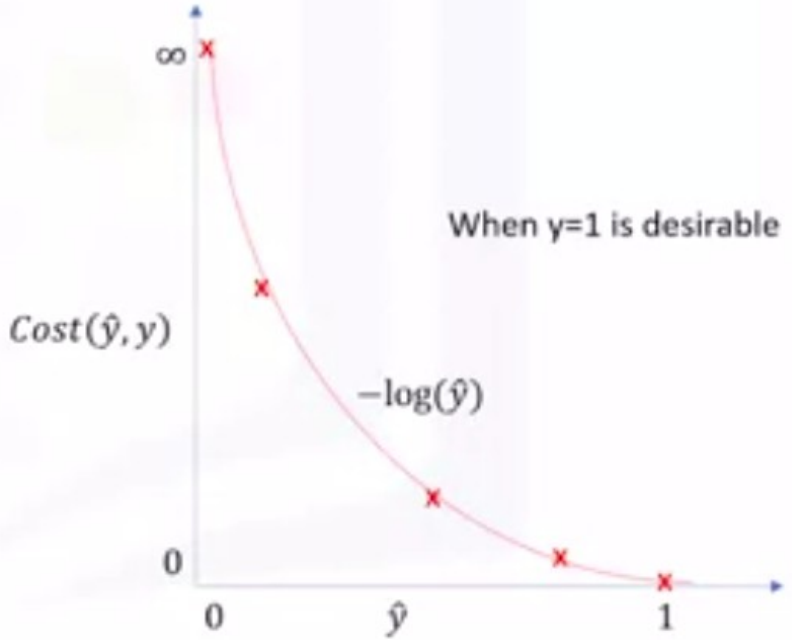
$$J(\theta) = \frac{1}{m}\sum_{i-1}^{m} Cost(\hat{y}, y)$$

Plotting the cost function of the model

- Model $\hat{y}$
- Actual Value y=1 or 0

- If Y=1, and $\hat{y}$=1 → cost = 0

- If Y=1, and $\hat{y}$=0 → cost = large



Logistic regression cost function

- So, we will replace cost function with:

$$Cost(\hat{y}, y) = \frac{1}{2}\left(\sigma(\theta^T X) - y\right)^2$$

$$Cost(\hat{y}, y) = \begin{cases} -\log(\hat{y}) & if\ y = 1 \\ -\log(1-\hat{y}) & if\ y = 0 \end{cases}$$

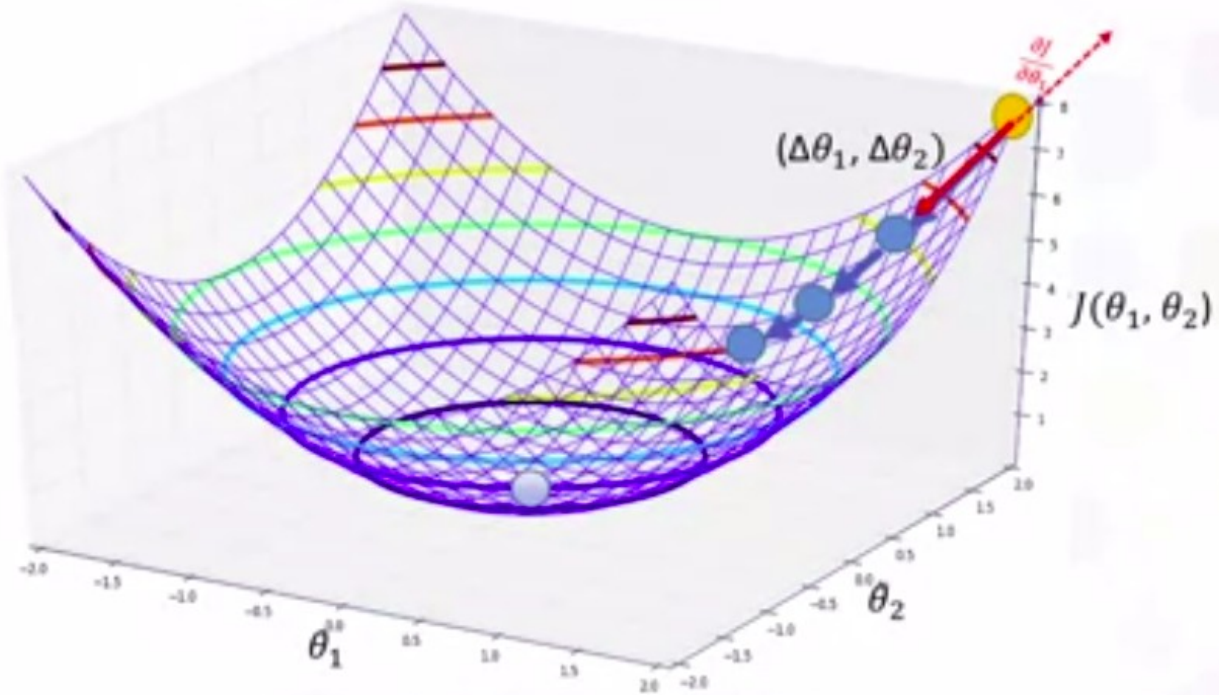
$$J(\theta) = \frac{1}{m}\sum_{i-1}^{m} Cost(\hat{y}, y)$$

$$J(\theta) = -\frac{1}{m}\sum_{i-1}^{m} y^i \log(\hat{y}^i) + (1-y^i)\log(1-\hat{y}^i)$$

Minimizing the cost function of the model

- How to find the best parameters for our model?
  - Minimize the cost function
- How to minimize the cost function?
  - Using Gradient Descent
- What is gradient descent?
  - A technique to use the derivative of a cost function to change the parameter values, in order to minimize the cost

Using gradient descent to minimize the cost



$$\frac{\partial J}{\partial \theta_1} = -\frac{1}{m}\sum_{i=1}^{m}\left(y^i - \hat{y}^i\right)x_1^i$$

$$\nabla J = \begin{bmatrix} \frac{\partial J}{\partial \theta_1} \\ \frac{\partial J}{\partial \theta_2} \\ \frac{\partial J}{\partial \theta_3} \\ \cdots \\ \cdots \\ \frac{\partial J}{\partial \theta_k} \end{bmatrix}$$

$$New\theta = old\theta - \eta\,\nabla J$$

$$\hat{y} = \sigma(\theta_1 x_1 + \theta_2 x_2)$$

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m} y^i \log\left(\hat{y}^i\right) + (1 - y^i)\log\left(1 - \hat{y}^i\right)$$

IBM Developer

SKILLS NETWORK

Question

What is "**gradient descent**" in training process?

- (●) A technique to use derivative of a cost function to change the parameter values, to minimize the cost.
- ( ) A technique to calculate the cost of logistic regression.
- ( ) A technique to initialize the parameters in training process.

✓ **Correct**

Skip Continue

Training algorithm recap

1. initialize the parameters randomly.
2. Feed the cost function with training set, and calculate the error.
3. Calculate the gradient of cost function.
4. Update weights with new values.
5. Go to step 2 until cost is small enough.
6. Predict the new customer X.

$$\theta^T = [\theta_0, \theta_1, \theta_2, \ldots.]$$

$$J(\theta) = -\frac{1}{m}\sum_{i-1}^{m} y^i \log(\hat{y}^i) + (1 - y^i)\log(1 - \hat{y}^i)$$

$$\nabla J = [\frac{\partial J}{\partial \theta_1}, \frac{\partial J}{\partial \theta_2}, \frac{\partial J}{\partial \theta_3}, \cdots, \frac{\partial J}{\partial \theta_k}]$$

$$\theta_{new} = \theta_{prv} - \eta \nabla J$$

$$\mathbf{P(y=1|x)} = \sigma(\theta^T X)$$

Support Vector Machine

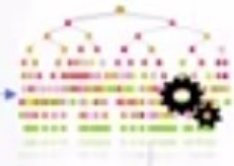
Classification with SVM

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000025	5	1	1	1	2	1	3	1	1	benign
1002945	5	4	4	5	7	10	3	2	1	benign
1015425	3	1	1	1	2	2	3	1	1	malignant
1016277	6	8	8	1	3	4	3	7	1	benign
1017023	4	1	1	3	2	1	3	1	1	benign
1017122	8	10	10	8	7	10		7	1	malignant
1018099	1	1	1	1	2	10	3	1	1	benign
1018561	2	1	2	H	2	1	3	1	1	benign
1033078	2	1	1	1	2	1	1	1	5	benign
1033078	4	2	1	1	2	1	2	1	1	benign

Modeling

ID	Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
1000015	6	1	1	1	7	1	3	1	1	Benign

Prediction



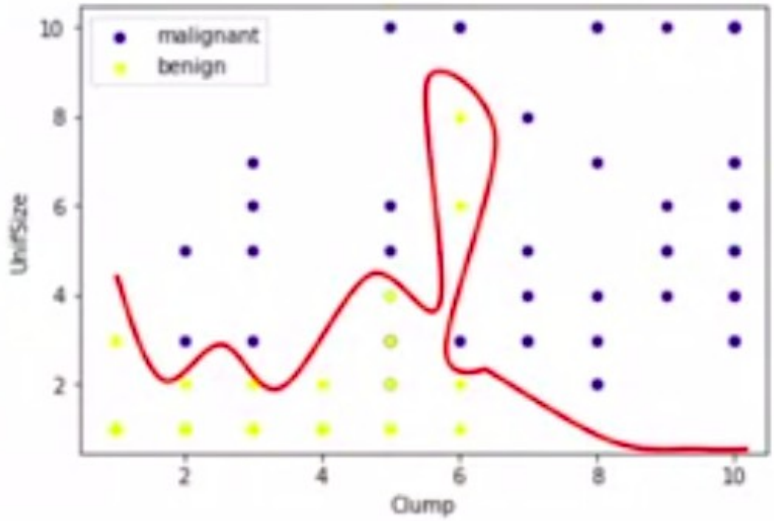
Accuracy = 89%

What is SVM?

SVM is a supervised algorithm that classifies cases by finding a separator.

1. Mapping data to a **high-dimensional** feature space
2. Finding a **separator**

Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
5	1	1	1	2	1	3	1	1	benign
5	4	4	5	7	10	3	2	1	benign
3	1	1	1	2	2	3	1	1	malignant
6	8	8	1	3	4	3	7	1	benign
4	1	1	3	2	1	3	1	1	benign
8	10	10	8	7	10		7	1	malignant
1	1	1	1	2	10	3	1	1	benign
2	1	2	H	2	1	3	1	1	benign
2	1	1	1	2	1	1	1	5	benign
4	2	1	1	2	1	2	1	1	benign

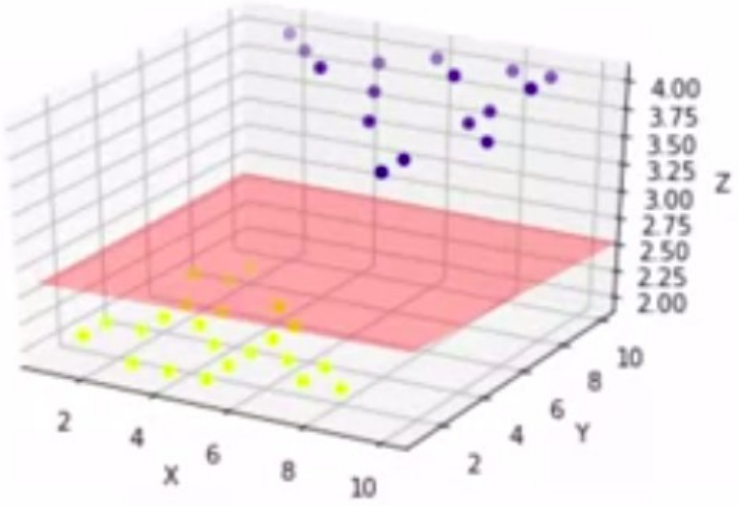


What is SVM?

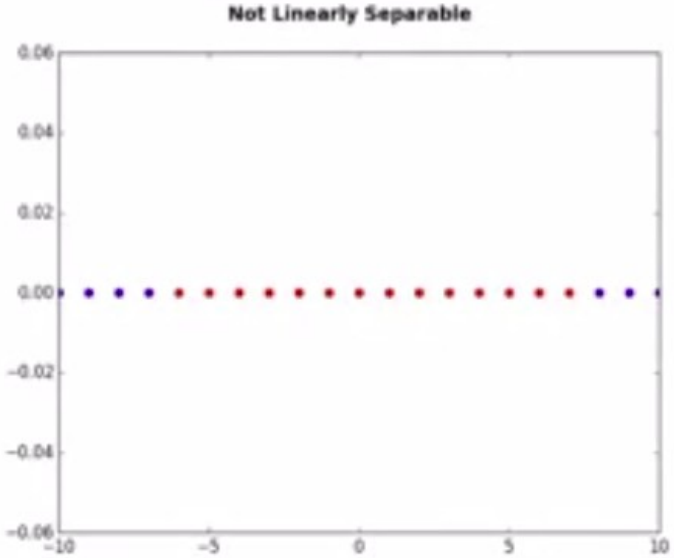
SVM is a supervised algorithm that classifies cases by finding a separator.

1. Mapping data to a **high-dimensional** feature space
2. Finding a **separator**

Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
5	1	1	1	2	1	3	1	1	benign
5	4	4	5	7	10	3	2	1	benign
3	1	1	1	2	2	3	1	1	malignant
6	8	8	1	3	4	3	7	1	benign
4	1	1	3	2	1	3	1	1	benign
8	10	10	8	7	10		7	1	malignant
1	1	1	1	2	10	3	1	1	benign
2	1	2	H	2	1	3	1	1	benign
2	1	1	1	2	1	1	1	5	benign
4	2	1	1	2	1	2	1	1	benign

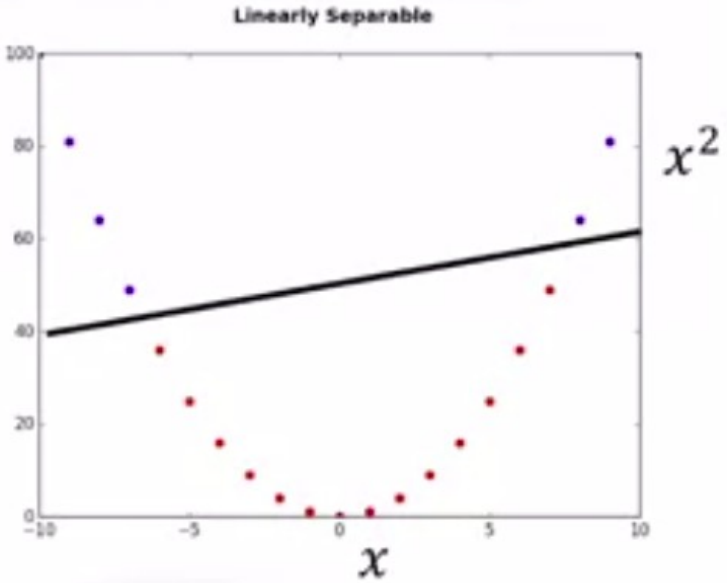


Data transformation



Kernelling:

- Linear
- Polynomial
- RBF
- Sigmoid



$$\phi(x) = [x, x^2]$$

Question

What is the meaning of "**Kernelling**" in SVM?

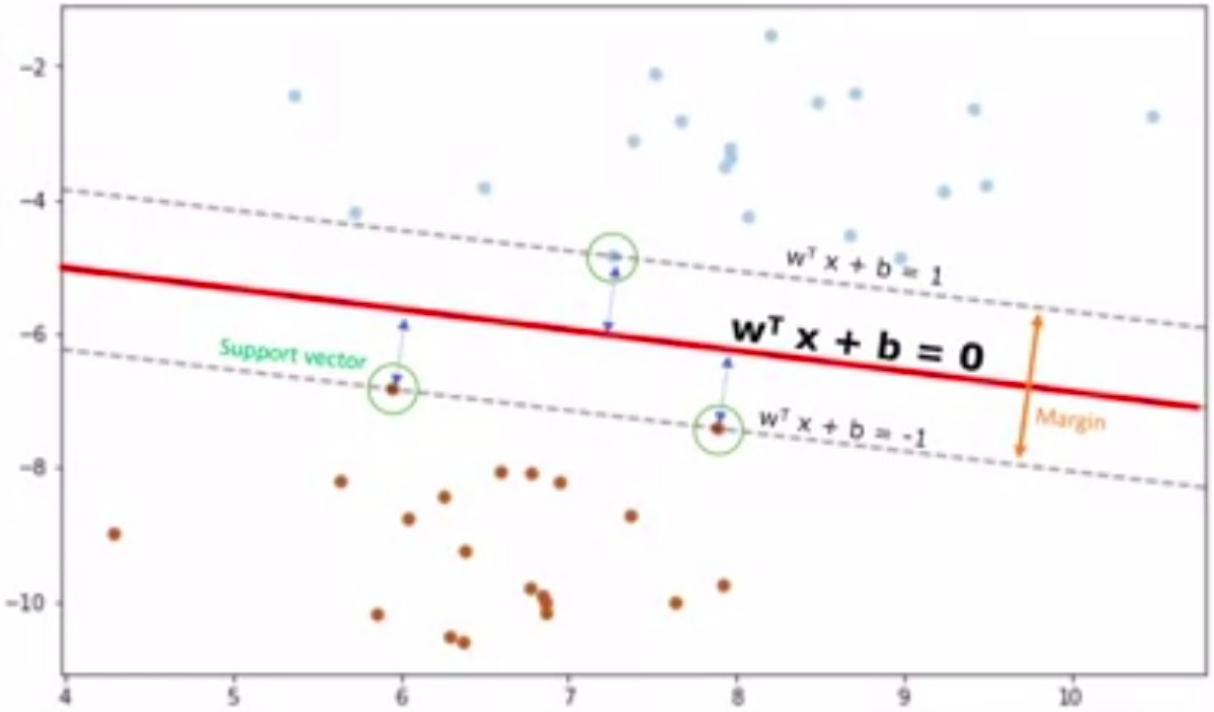
- (●) Mapping data into a higher dimensional space, in such a way that can change a linearly inseparable dataset into a linearly separable dataset.
- ( ) A function to reduce the dimensionality of a dataset in SVM.
- ( ) Finding a hyperplane in such a way that increase the dimensionality of a dataset.

✓ **Correct**

Skip Continue

Using SVM to find the hyperplane

Find $\mathbf{w}$ and b such that
$\Phi(\mathbf{w}) = \frac{1}{2}\,\mathbf{w}^T\mathbf{w}$ is minimized;
and for all $\{(\mathbf{x}_i, y_i)\}$: $y_i\,(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$



Pros and cons of SVM

- Advantages:
  - Accurate in high-dimensional spaces
  - Memory efficient

- Disadvantages:
  - Prone to over-fitting
  - No probability estimation
  - Small datasets

SVM applications

- Image recognition
- Text category assignment
- Detecting spam
- Sentiment analysis
- Gene Expression Classification
- Regression, outlier detection and clustering

Back

Practice Quiz: Linear Classification

Practice Quiz • 10 min • 3 total points

1. Which of the following examples is/are a sample application of Logistic Regression? (select three)

1 / 1 point

- [x] The probability that a person has a heart attack within a specified time period using person's age and sex.

Correct
Correct! The outcome is binary and uses other variables as predictors.

- [ ] Estimating the blood pressure of a patient based on her symptoms and biographical data.
- [x] Likelihood of a homeowner defaulting on a mortgage.

Correct
Correct! Here, we try to predict the possibility of defaulting versus not defaulting, which is a categorical response.

- [x] Customer's propensity to purchase a product or halt a subscription in marketing applications.

Correct
Correct! The outcome is a probability of a categorical variable.

Practice Quiz: Linear Classification

Practice Quiz • 10 min • 3 total points

Correct

Correct! The outcome is a probability of a categorical variable.

2. Which of the following statements comparing linear and logistic regressions is TRUE?

1 / 1 point

- (●) Linear regression is used for a continuous target whereas logistic regression is more suitable for a categorical target.
- ( ) Independent variables in linear regression can be continuous or categorical, but can only be categorical in logistic regression.
- ( ) Both linear and logistic regression can be used to predict categorical responses and attain a point's likelihood of belonging to each class.
- ( ) In this course, linear regression minimizes the mean absolute error, while logistic regression minimizes the mean squared error.

Correct

Correct! Linear regression is not suitable for a categorical target because it tries to fit a line through the data, but the prediction is a step function that doesn't reflect class probability well.

Correct
Correct! Linear regression is not suitable for a categorical target because it tries to fit a line through the data, but the prediction is a step function that doesn't reflect class probability well.

3. How are gradient descent and learning rate used in logistic regression? 1 / 1 point

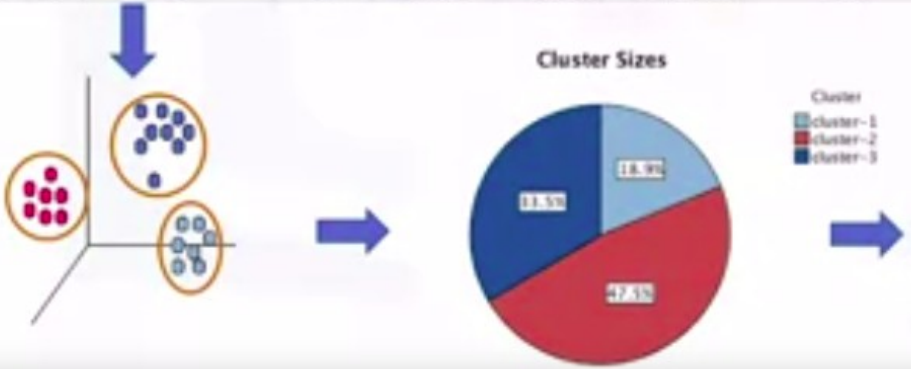
- ○ Gradient descent takes increasingly bigger steps towards the minimum with each iteration.
- ◉ Gradient descent specifies the steps to take in the current slope direction, learning rate is the step length.
- ○ We want to minimize the cost by maximizing the learning rate value.
- ○ Gradient descent will minimize learning rate to minimize the cost in fewer iterations.

Correct
Correct! Gradient descent takes steps toward the minimum of the cost function, and the learning rate gives us control over how fast we move.

Intro to Clustering

Clustering for segmentation

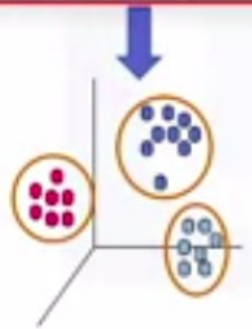
Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

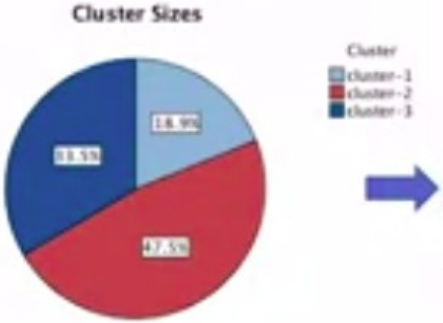


Cluster	Segment Name
cluster-1	AFFULUENT AND MIDDLE AGED
cluster-2	YOUNG EDUCATED AND MIDDLE INCOME
cluster-3	YOUNG AND LOW INCOME

Clustering for segmentation

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1





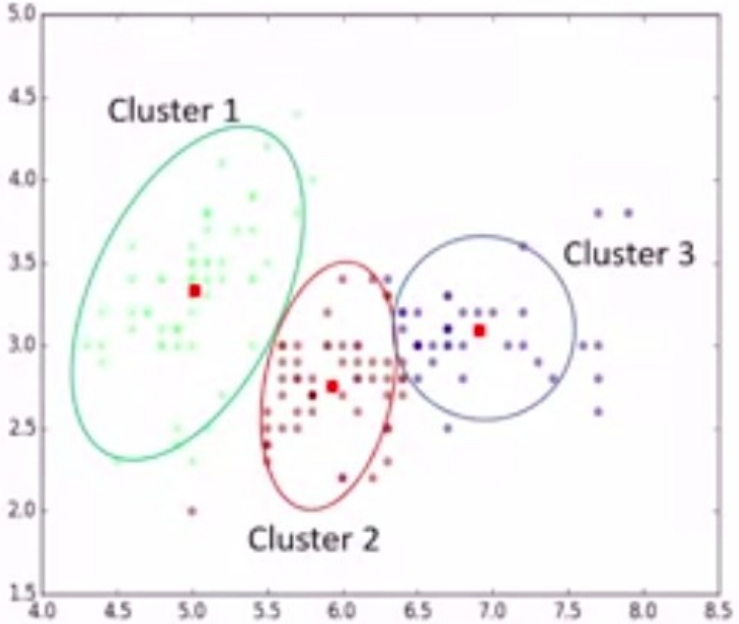
Customer ID	Segment
1	YOUNG AND LOW INCOME
2	AFFULUENT AND MIDDLE AGED
3	AFFULUENT AND MIDDLE AGED
4	YOUNG AND LOW INCOME
5	AFFULUENT AND MIDDLE AGED
6	AFFULUENT AND MIDDLE AGED
7	YOUNG AND LOW INCOME
8	YOUNG AND LOW INCOME
9	AFFULUENT AND MIDDLE AGED

Cluster	Segment Name
cluster-1	AFFULUENT AND MIDDLE AGED
cluster-2	YOUNG EDUCATED AND MIDDLE INCOME
cluster-3	YOUNG AND LOW INCOME

What is clustering?

What is a cluster?

A group of objects that are **similar to other objects** in the cluster, and **dissimilar to data points** in other clusters.



Clustering Vs. classification

Labeled dataset

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1
10	47	3	23	115	0.653	3.947	NBA011	4	0

Modeling

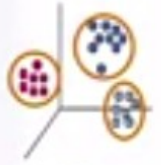
Decision Tree

Prediction

Unlabeled dataset

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

Modeling

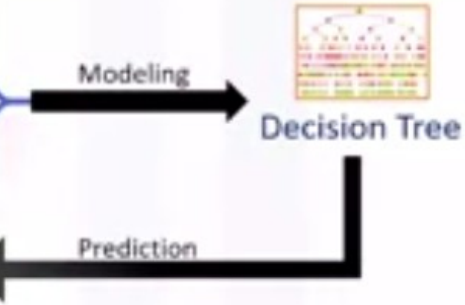


K-Means

Clustering Vs. classification

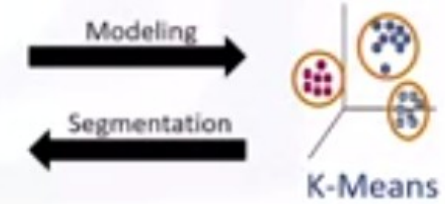
Labeled dataset

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1
10	47	3	23	115	0.653	3.947	NBA011	4	0



Unlabeled dataset

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1



Question

"Clustering algorithms predict categorical class labels," is it TRUE or FALSE?

- ○ TRUE
- ◉ FALSE

✓ **Correct**

Skip | Continue

Clustering applications

- RETAIL/MARKETING:
  - Identifying buying patterns of customers
  - Recommending new books or movies to new customers
- BANKING:
  - Fraud detection in credit card use
  - Identifying clusters of customers (e.g., loyal)
- INSURANCE:
  - Fraud detection in claims analysis
  - Insurance risk of customers

Clustering applications

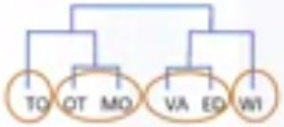
- PUBLICATION:
  - Auto-categorizing news based on their content
  - Recommending similar news articles
- MEDICINE:
  - Characterizing patient behavior
- BIOLOGY:
  - Clustering genetic markers to identify family ties

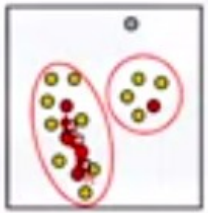
Why clustering?

- Exploratory data analysis
- Summary generation
- Outlier detection
- Finding duplicates
- Pre-processing step

Clustering algorithms

- Partitioned-based Clustering
  - Relatively efficient
  - E.g. k-Means, k-Median, Fuzzy c-Means
- Hierarchical Clustering
  - Produces trees of clusters
  - E.g. Agglomerative, Divisive
- Density-based Clustering
  - Produces arbitrary shaped clusters
  - E.g. DBSCAN

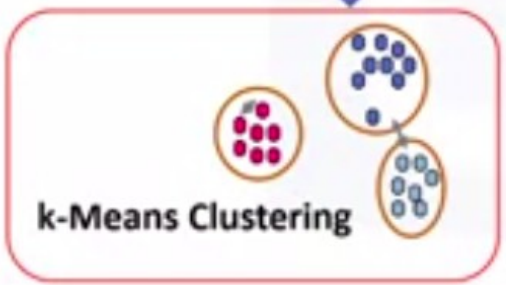


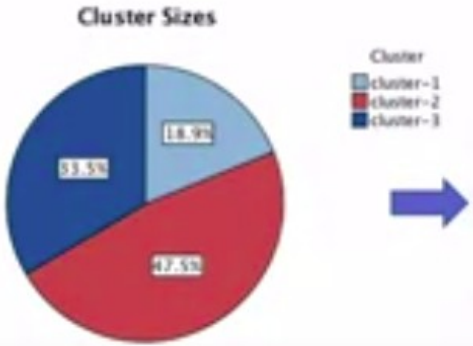


What is k-Means clustering?

Customer Id	Age	Edu	Years Employed	Income	Card Debt	Other Debt	Address	DebtIncomeRatio	Defaulted
1	41	2	6	19	0.124	1.073	NBA001	6.3	0
2	47	1	26	100	4.582	8.218	NBA021	12.8	0
3	33	2	10	57	6.111	5.802	NBA013	20.9	1
4	29	2	4	19	0.681	0.516	NBA009	6.3	0
5	47	1	31	253	9.308	8.908	NBA008	7.2	0
6	40	1	23	81	0.998	7.831	NBA016	10.9	1
7	38	2	4	56	0.442	0.454	NBA013	1.6	0
8	42	3	0	64	0.279	3.945	NBA009	6.6	0
9	26	1	5	18	0.575	2.215	NBA006	15.5	1

Customer ID	Segment
1	YOUNG AND LOW INCOME
2	AFFULUENT AND MIDDLE AGED
3	AFFULUENT AND MIDDLE AGED
4	YOUNG AND LOW INCOME
5	AFFULUENT AND MIDDLE AGED
6	AFFULUENT AND MIDDLE AGED
7	YOUNG AND LOW INCOME
8	YOUNG AND LOW INCOME
9	AFFULUENT AND MIDDLE AGED

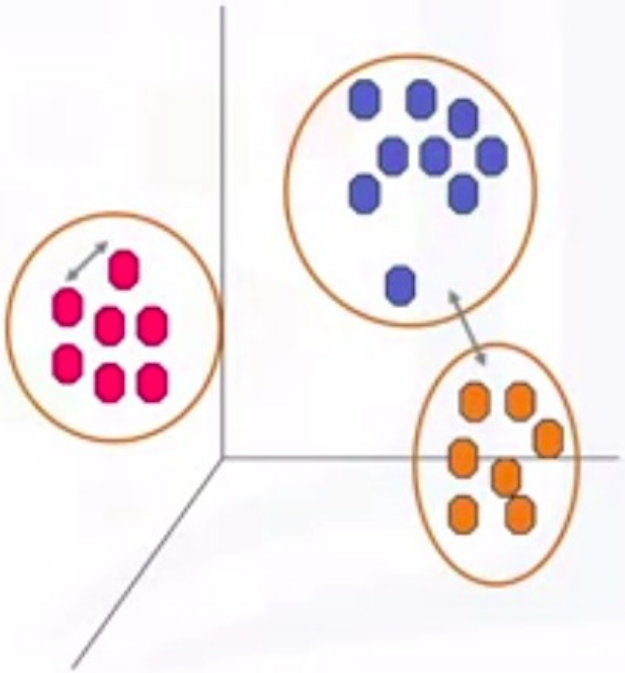




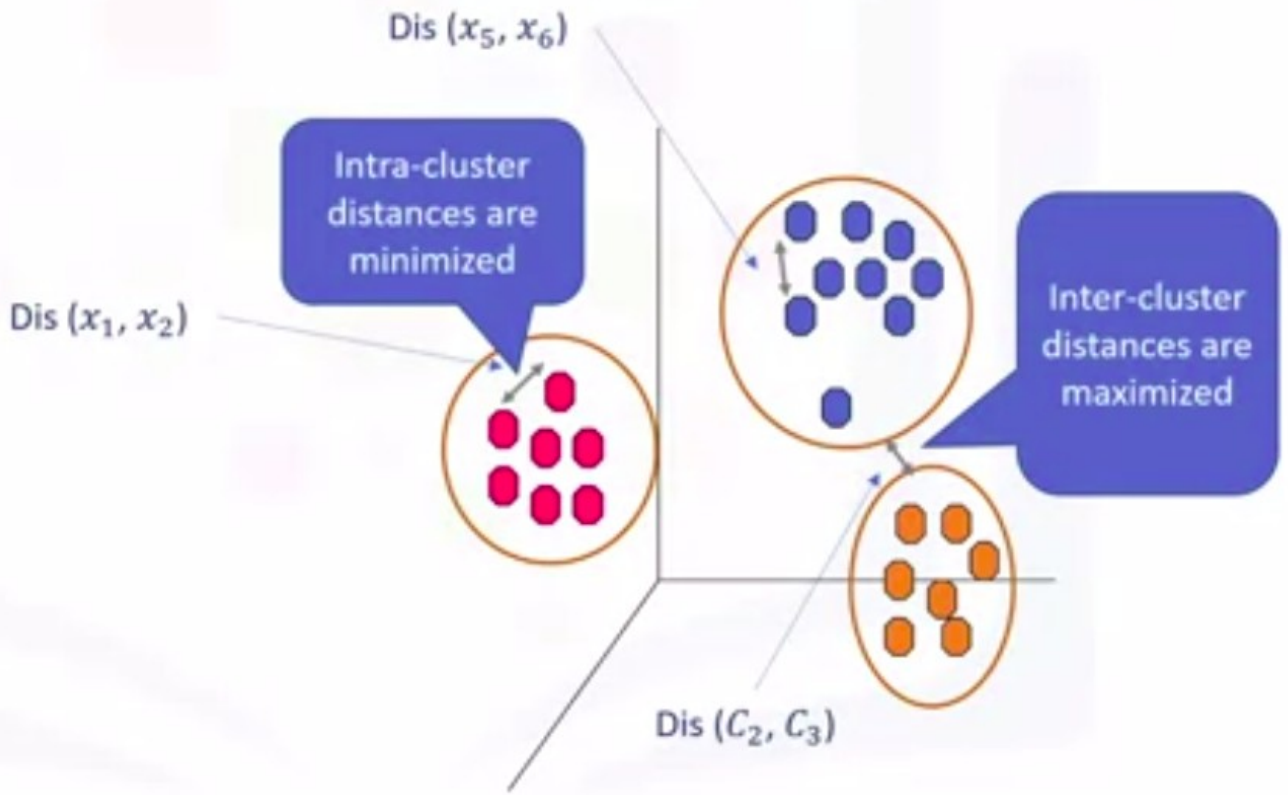
Cluster	Segment Name
cluster-1	AFFULUENT AND MIDDLE AGED
cluster-2	YOUNG EDUCATED AND MIDDLE INCOME
cluster-3	YOUNG AND LOW INCOME

k-Means algorithms

- Partitioning Clustering
- K-means divides the data into non-overlapping subsets (clusters) without any cluster-internal structure
- Examples within a cluster are very similar
- Examples across different clusters are very different



Determine the similarity or dissimilarity



Question

What is the objective of k-means?

- [x] To form clusters in such a way that similar samples go into a cluster, and dissimilar samples fall into different clusters.

  Correct

- [x] To minimize the “intra cluster” distances and maximize the “inter-cluster” distances.

  Correct

- [x] To divide the data into non-overlapping clusters without any cluster-internal structure

  Correct

Skip Continue

1-dimensional similarity/distance

Customer 1
Age
54

Customer 2
Age
50

$$\text{Dis}\,(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

$$\text{Dis}\,(x_1, x_2) = \sqrt{(54 - 50)^2} = 4$$

Multi-dimentional similarity/distance

Customer 1		
Age	Income	education
54	190	3

Customer 2		
Age	Income	education
50	200	8

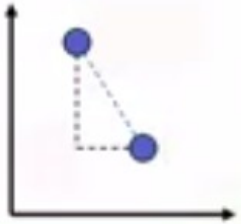
$$\text{Dis}\,(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

$$= \sqrt{(54-50)^2+(190-200)^2+(3-8)^2} = 11.87$$

2-dimensional similarity/distance

Customer 1

Age	Income
54	190



Customer 2

Age	Income
50	200

$$\text{Dis}\,(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

$$= \sqrt{(54-50)^2+(190-200)^2} = 10.77$$

Multi-dimentional similarity/distance

Customer 1		
Age	Income	education
54	190	3

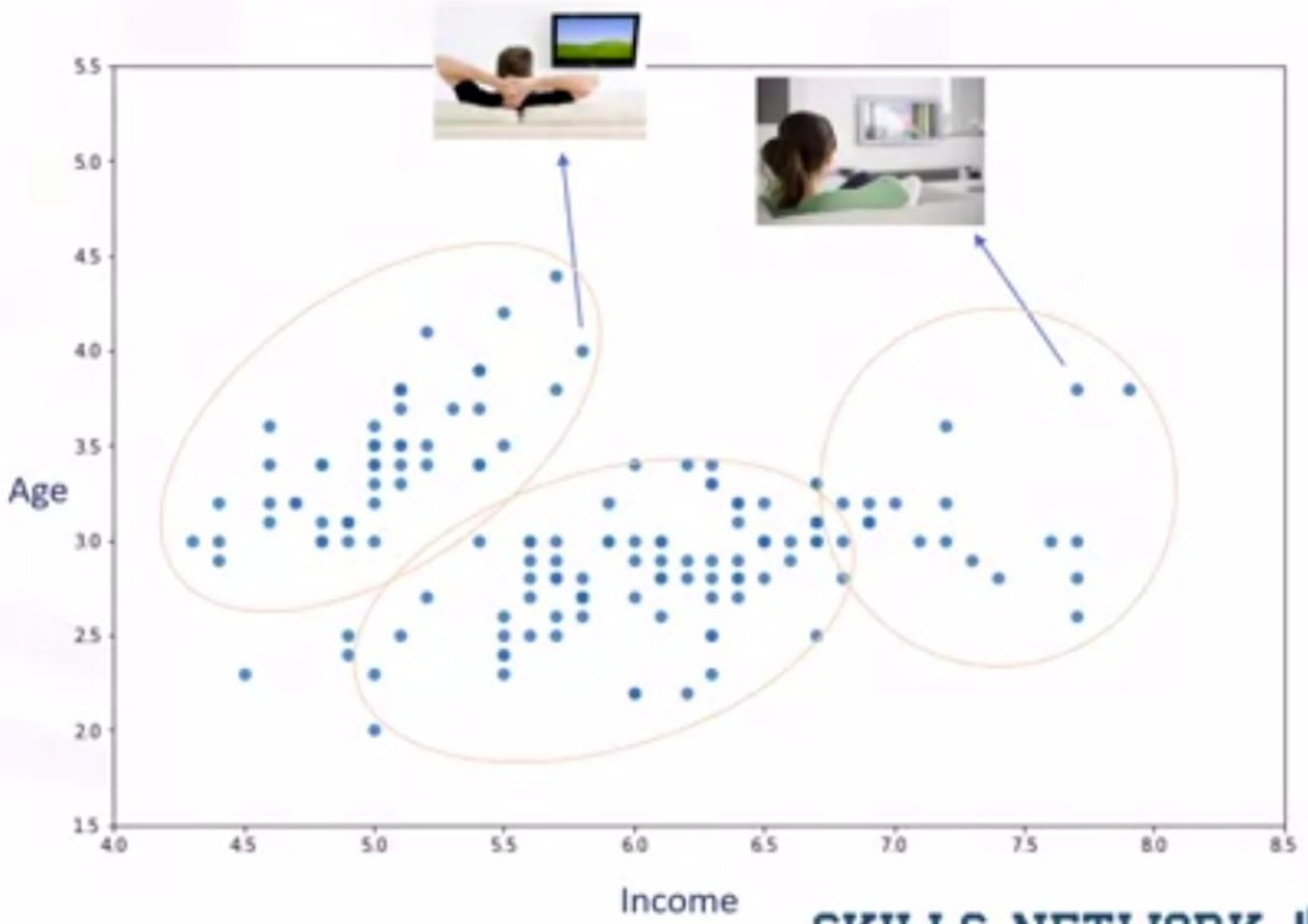
Customer 2		
Age	Income	education
50	200	8

$$\text{Dis}\,(x_1, x_2) = \sqrt{\sum_{i=0}^{n}(x_{1i} - x_{2i})^2}$$

$$= \sqrt{(54-50)^2+(190-200)^2+(3-8)^2} = 11.87$$

How does k-Means clustering work?

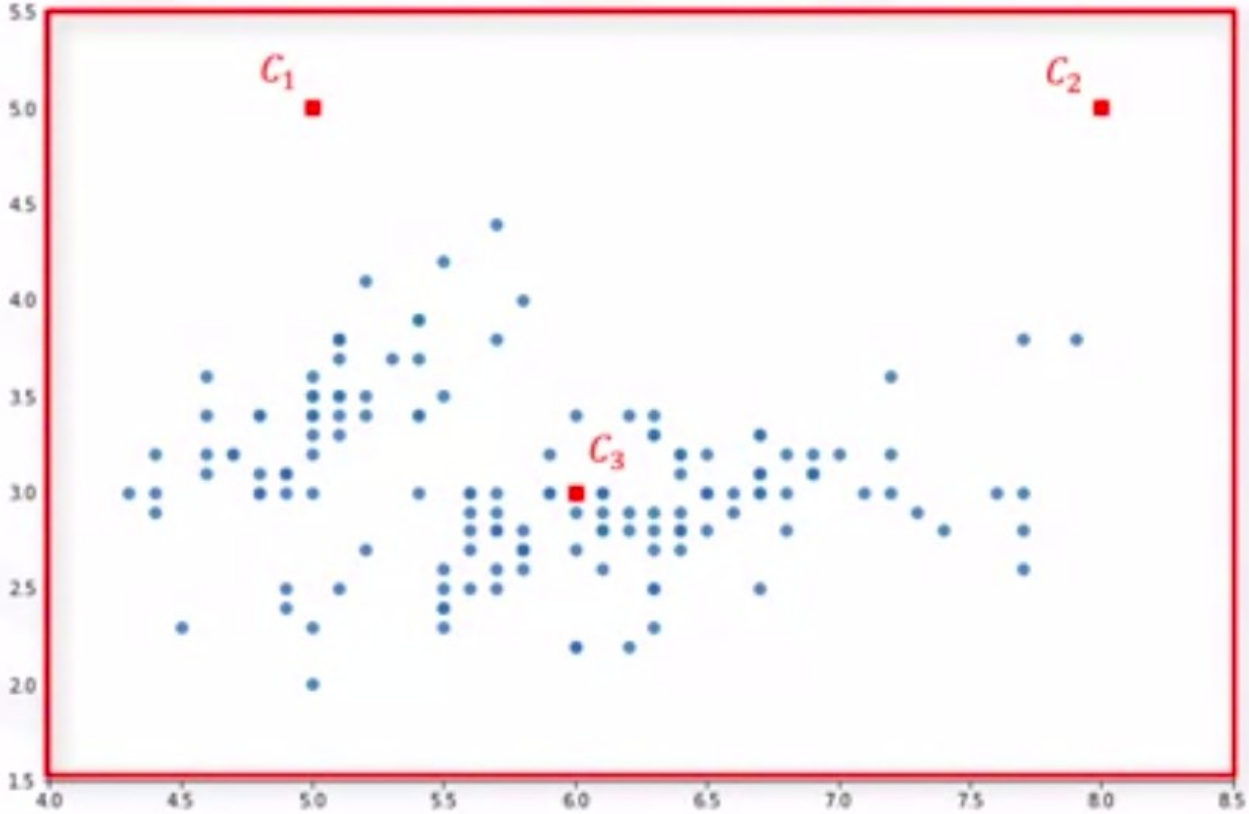
Customer ID	Age	Income
1	3	4
2	2	6
3	3.5	2
...	...	..



k-Means clustering – initialize k

1) Initialize k=3 centroids randomly

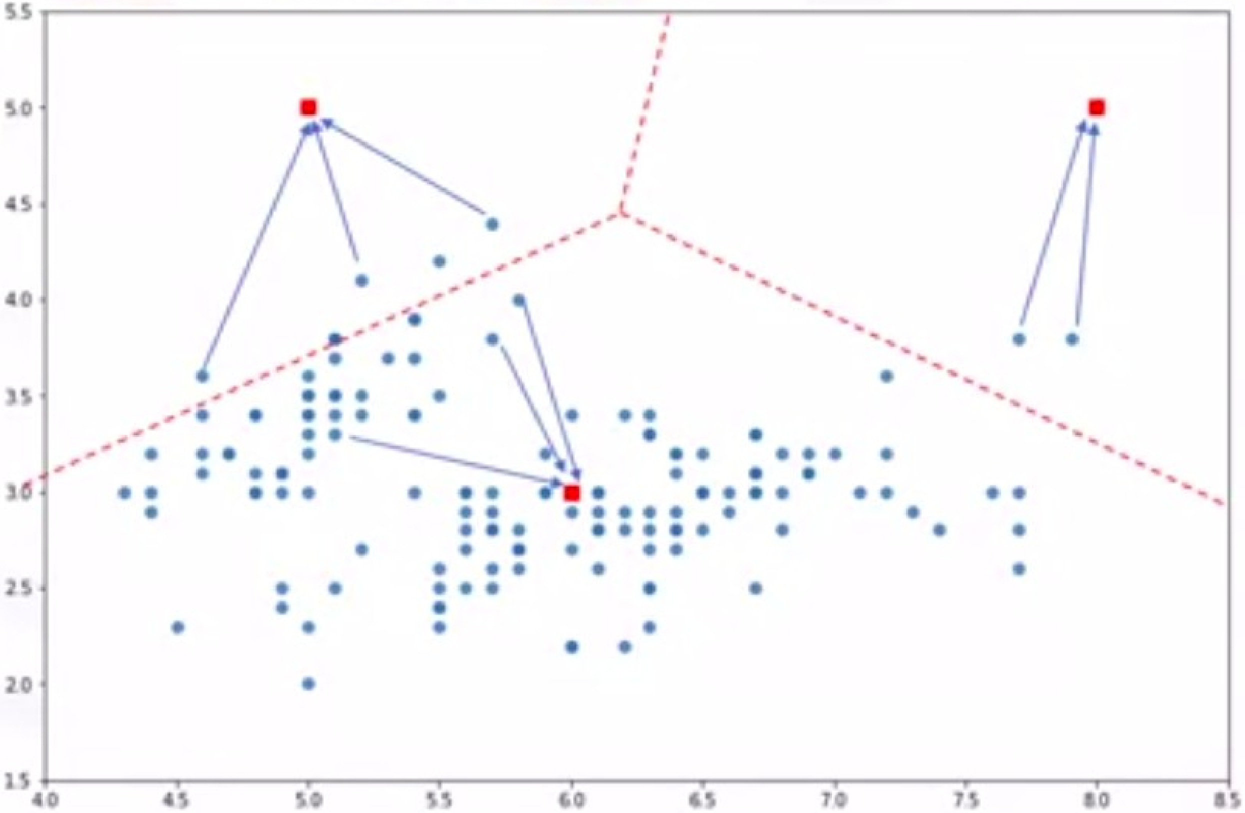
C_1 = [8., 5.]
C_2 = [5., 5.]
C_3 = [6., 3.]



k-Means clustering – assign to centroid

3) Assign each point to the closest centroid

C_1	C_2	C_3
d(p1, c1)	d(p1, c2)	d(p1, c3)
d(p2, c1)	d(p2, c2)	d(p2, c3)
d(p3, c1)	d(p3, c2)	d(p3, c3)
d(p4, c1)	d(p4, c2)	d(p4, c3)
d(p ..., c1)	d(p.., c2)	d(p.., c3)
d(pn, c1)	d(pn, c2)	d(pn, c3)
d(p.., c1)	d(p.., c2)	d(p.., c3)
d(p ..., c1)	d(p.., c2)	d(p.., c3)
d(pn, c1)	d(pn, c2)	d(pn, c3)
d(p.., c1)	d(p.., c2)	d(p.., c3)
d(p ..., c1)	d(p.., c2)	d(p.., c3)
d(pn, c1)	d(pn, c2)	d(pn, c3)
d(p.., c1)	d(p.., c2)	d(p.., c3)
d(p ..., c1)	d(p.., c2)	d(p.., c3)
d(pn, c1)	d(pn, c2)	d(pn, c3)

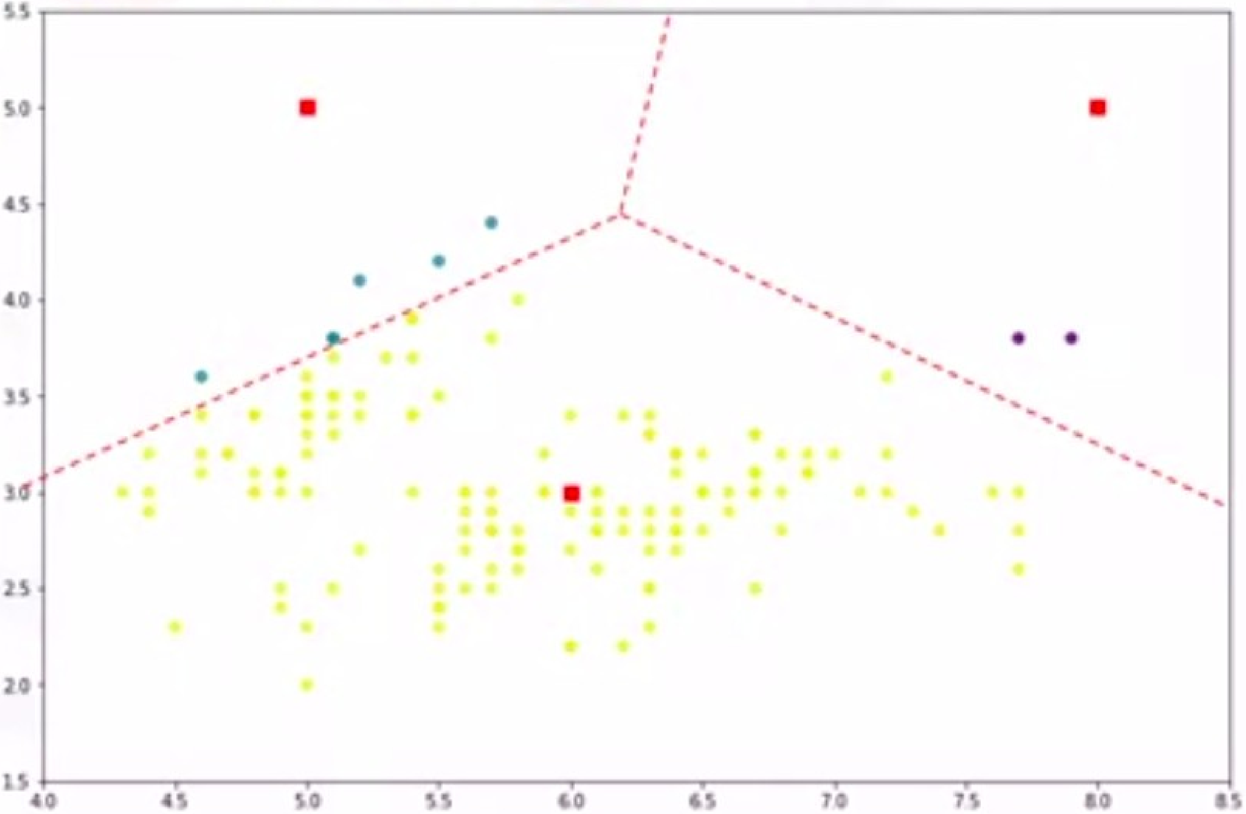


k-Means clustering – assign to centroid

3) Assign each point to the closest centroid

$$
\begin{array}{ccc}
C_1 & C_2 & C_3
\end{array}
$$

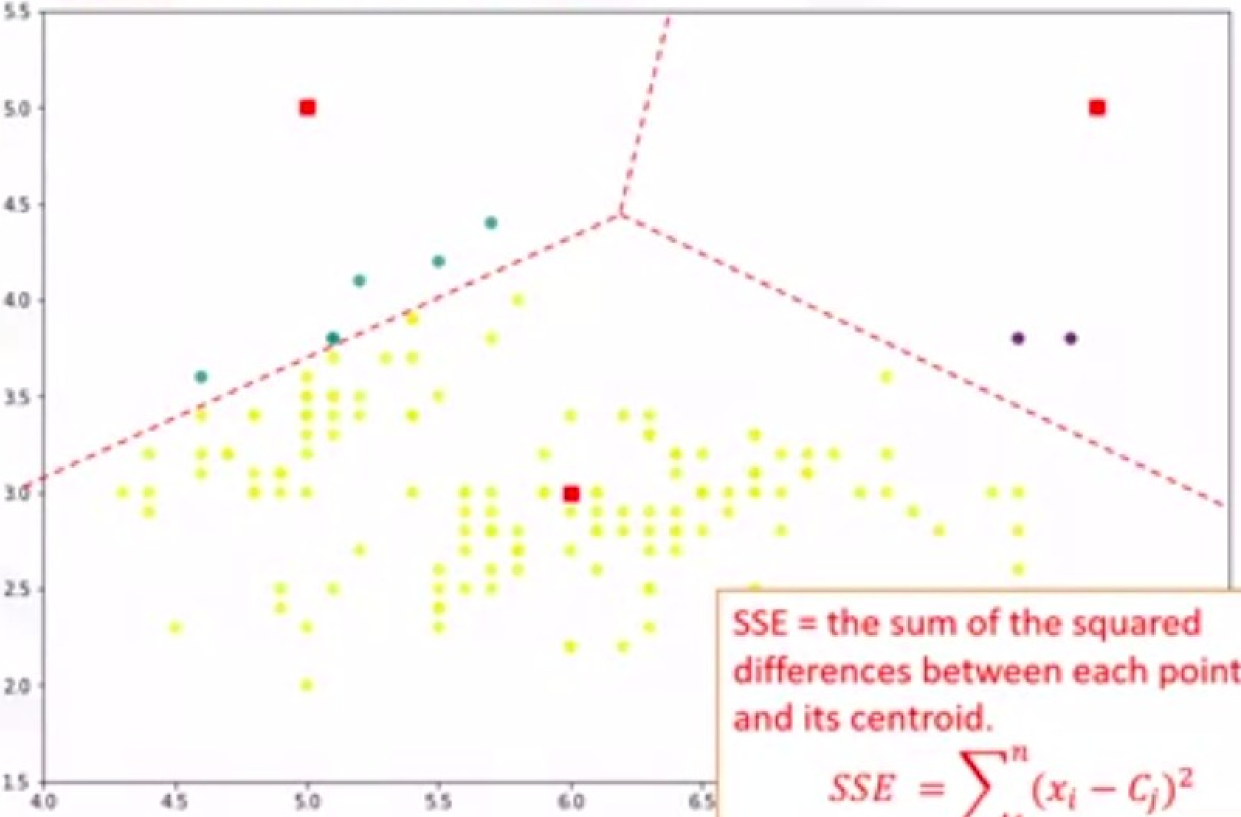
$$
\begin{bmatrix}
d(p1,c1) & d(p1,c2) & d(p1,c3) \\
d(p2,c1) & d(p2,c2) & d(p2,c3) \\
d(p3,c1) & d(p3,c2) & d(p3,c3) \\
d(p4,c1) & d(p4,c2) & d(p4,c3) \\
d(p\ldots,c1) & d(p..,c2) & d(p..,c3) \\
d(pn,c1) & d(pn,c2) & d(pn,c3) \\
d(p..,c1) & d(p..,c2) & d(p..,c3) \\
d(p\ldots,c1) & d(p..,c2) & d(p..,c3) \\
d(pn,c1) & d(pn,c2) & d(pn,c3) \\
d(p..,c1) & d(p..,c2) & d(p..,c3) \\
d(p\ldots,c1) & d(p..,c2) & d(p..,c3) \\
d(pn,c1) & d(pn,c2) & d(pn,c3) \\
d(p..,c1) & d(p..,c2) & d(p..,c3) \\
d(p\ldots,c1) & d(p..,c2) & d(p..,c3) \\
d(pn,c1) & d(pn,c2) & d(pn,c3)
\end{bmatrix}
$$



k-Means clustering – assign to centroid

3) Assign each point to the closest centroid

C_1	C_2	C_3
$d(p1,c1)$	$d(p1,c2)$	$d(p1,c3)$
$d(p2,c1)$	$d(p2,c2)$	$d(p2,c3)$
$d(p3,c1)$	$d(p3,c2)$	$d(p3,c3)$
$d(p4,c1)$	$d(p4,c2)$	$d(p4,c3)$
$d(p\ ...,c1)$	$d(p..,c2)$	$d(p..,c3)$
$d(pn,c1)$	$d(pn,c2)$	$d(pn,c3)$
$d(p..,c1)$	$d(p..,c2)$	$d(p..,c3)$
$d(p\ ...,c1)$	$d(p..,c2)$	$d(p..,c3)$
$d(pn,c1)$	$d(pn,c2)$	$d(pn,c3)$
$d(p..,c1)$	$d(p..,c2)$	$d(p..,c3)$
$d(p\ ...,c1)$	$d(p..,c2)$	$d(p..,c3)$
$d(pn,c1)$	$d(pn,c2)$	$d(pn,c3)$
$d(p..,c1)$	$d(p..,c2)$	$d(p..,c3)$
$d(p\ ...,c1)$	$d(p..,c2)$	$d(p..,c3)$
$d(pn,c1)$	$d(pn,c2)$	$d(pn,c3)$



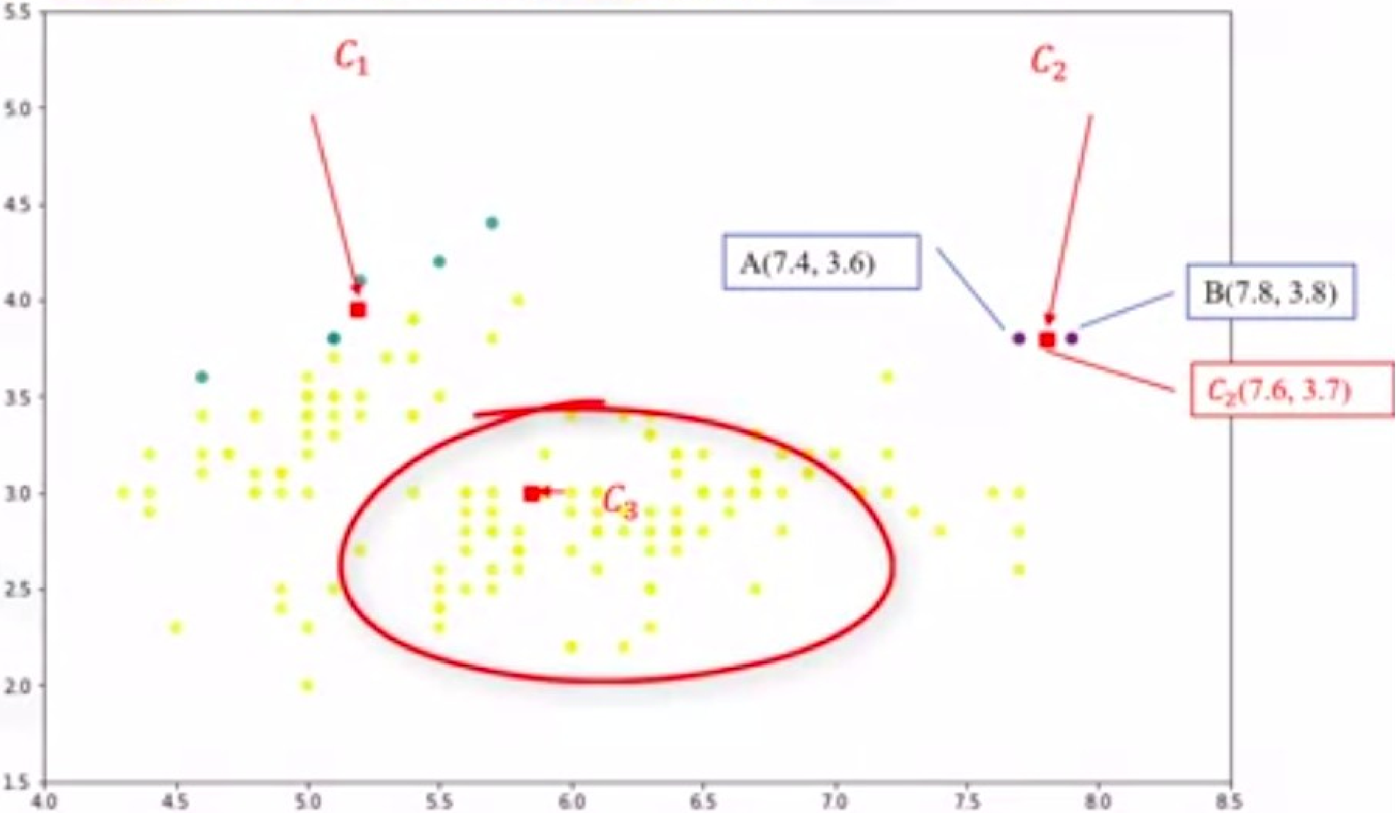
SSE = the sum of the squared differences between each point and its centroid.

$$SSE = \sum_{1}^{n}(x_i - C_j)^2$$

k-Means clustering – compute new centroids

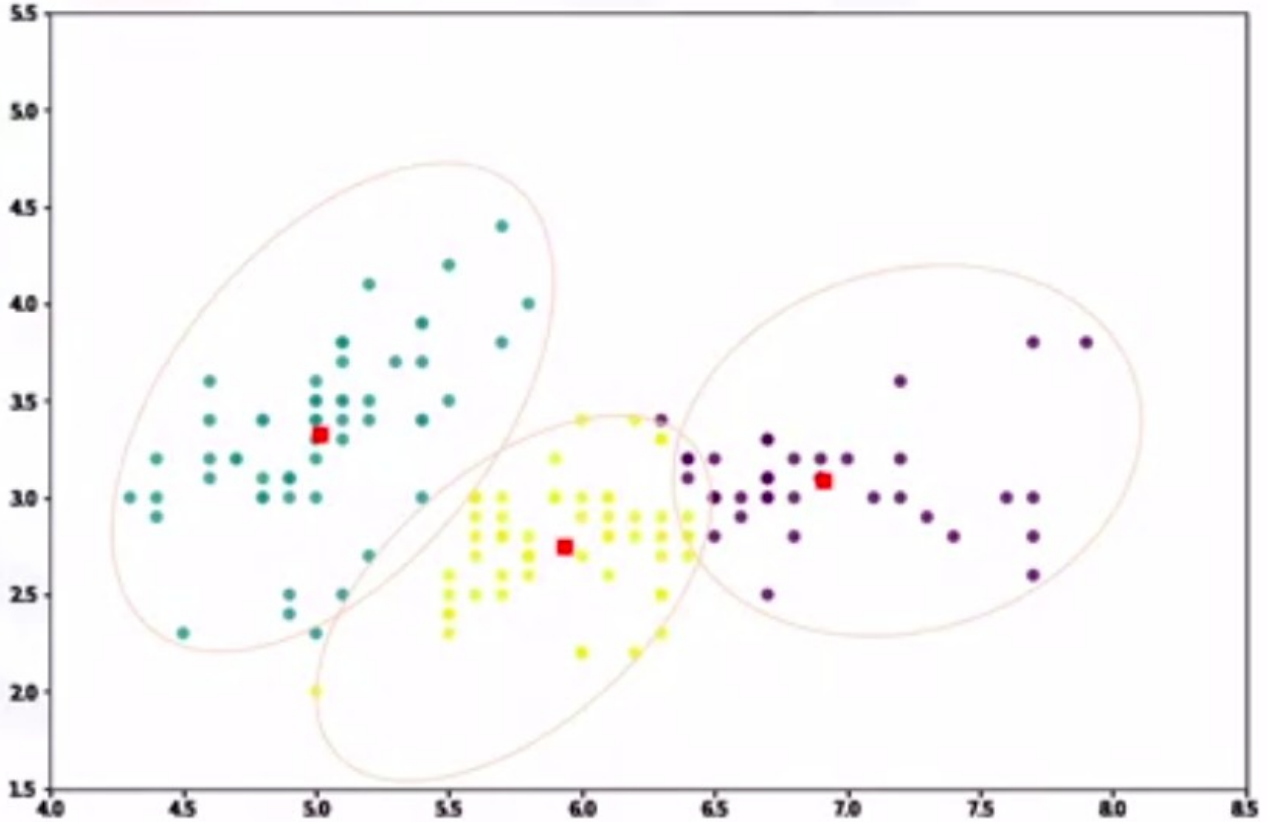
4) Compute the new centroids for each cluster.

C_1	C_2	C_3
d(p1, c1)	d(p1, c2)	d(p1, c3)
d(p2, c1)	d(p2, c2)	d(p2, c3)
d(p3, c1)	d(p3, c2)	d(p3, c3)
d(p4, c1)	d(p4, c2)	d(p4, c3)
d(p ..., c1)	d(p.., c2)	d(p.., c3)
d(pn, c1)	d(pn, c2)	d(pn, c3)
d(p.., c1)	d(p.., c2)	d(p.., c3)
d(p ..., c1)	d(p.., c2)	d(p.., c3)
d(pn, c1)	d(pn, c2)	d(pn, c3)
d(p.., c1)	d(p.., c2)	d(p.., c3)
d(p ..., c1)	d(p.., c2)	d(p.., c3)
d(pn, c1)	d(pn, c2)	d(pn, c3)
d(p.., c1)	d(p.., c2)	d(p.., c3)
d(p ..., c1)	d(p.., c2)	d(p.., c3)
d(pn, c1)	d(pn, c2)	d(pn, c3)



k-Means clustering – repeat

5) Repeat until there are no more changes.



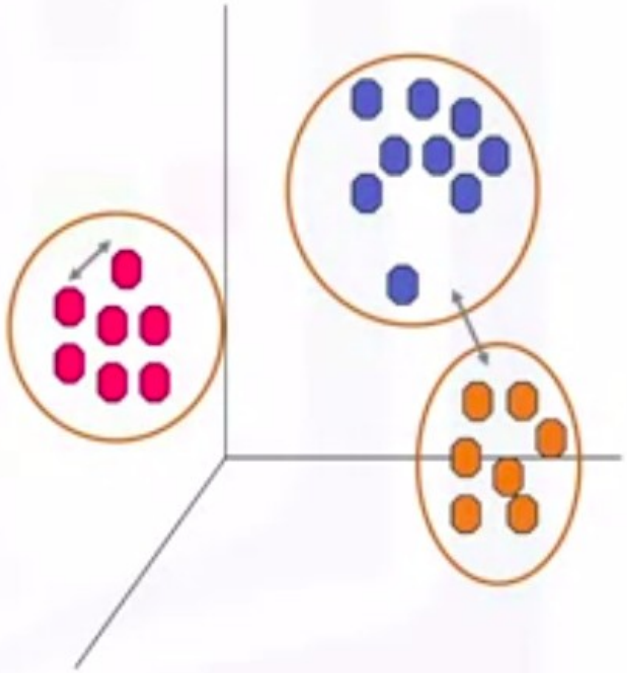
More on k-Means

k-Means clustering algorithm

1. Randomly placing *k* centroids, one for each cluster.
2. Calculate the distance of each point from each centroid.
3. Assign each data point (object) to its closest centroid, creating a cluster.
4. Recalculate the position of the *k* centroids.
5. Repeat the steps 2-4, until the centroids no longer move.

k-Means accuracy

- External approach
  - Compare the clusters with the ground truth, if it is available.
- Internal approach
  - Average the distance between data points within a cluster.



Choosing k


mean distance of data points to cluster centroid
elbow point, k=5
K=1
K=2
K=3
K=4
K=5
K=6
K=7
K=8
K=9

Question

In clustering evaluation process, "**elbow point**" is where the rate of accuracy increase sharply, when we run clustering multiple times, increasing k in each run.

- ○ TRUE
- ◉ FALSE

✓ **Correct**

Skip | Continue

k-Means recap

- Med and Large sized databases (*Relatively efficient*)
- Produces sphere-like clusters
- Needs number of clusters (k)

Back

Practice Quiz: Clustering

Practice Quiz • 10 min • 3 total points

Congratulations! You passed!

Grade received 100% **To pass** 66% or higher

Go to next item

1. Which of the following is an application of clustering? **1 / 1 point**

- ○ Customer churn prediction
- ○ Price estimation
- ○ Sales prediction
- ◉ Customer segmentation

Correct
Correct! Clustering can help partition individuals into groups with similar characteristics.

2. Which approach can be used to calculate dissimilarity of objects in clustering? **1 / 1 point**

- ○ Cosine similarity

Customer segmentation

Correct
Correct! Clustering can help partition individuals into groups with similar characteristics.

2. Which approach can be used to calculate dissimilarity of objects in clustering? **1 / 1 point**

- Cosine similarity
- Minkowski distance
- Euclidian distance
- All of the above

Correct
Correct! All of the approaches are valid approaches to calculate dissimilarity.

3. How is a center point (centroid) picked for each cluster in k-means upon initialization? (select two) **1 / 1 point**

- [ ] We select the k points closest to the mean/median of the entire dataset.
- [x] We can randomly choose some observations out of the data set and use these observations as the initial means.

Back

Practice Quiz: Clustering

Practice Quiz • 10 min • 3 total points

- (●) All of the above

Correct
Correct! All of the approaches are valid approaches to calculate dissimilarity.

3. How is a center point (centroid) picked for each cluster in k-means upon initialization? (select two)

1 / 1 point

- [ ] We select the k points closest to the mean/median of the entire dataset.
- [x] We can randomly choose some observations out of the data set and use these observations as the initial means.

Correct
Correct! These centroids will be updated based on the clusters formed at each iteration.

- [x] We can create some random points as centroids of the clusters.

Correct
Correct! We can randomly place k centroids, one for each cluster. Each data point is then assigned to its closest centroid.

- [ ] We can select it through correlation analysis.